Performance Tests

Anthony Wetherell

Human Factors Section, Chemical and Biological Defence Establishment, Porton Down, Salisbury, United Kingdom

This paper discusses the use of psychological performance tests to assess the effects of environmental stressors. The large number and the variety of performance tests are illustrated, and the differences between performance tests and other psychological tests are described in terms of their design, construction, use, and purpose. The stressor emphasis is on the effects of drugs since that is where most performance tests have found their main application, although other stressors, e.g., fatigue, toxic chemicals, are mentioned where appropriate. Diazepam is used as an example. There is no particular performance emphasis since the tests are intended to have wide applicability. However, vehicle-driving performance is discussed because it has been the subject of a great deal of research and is probably one of the most important areas of application. Performance tests are discussed in terms of the four main underlying models—factor analysis, general information processing, multiple resource and strategy models, and processing-stage models—and in terms of their psychometric properties—sensitivity, reliability, and content, criterion, construct, and face validity. Some test taxonomies are presented. Standardization is also discussed with reference to the reaction time, mathematical processing, memory search, spatial processing, unstable tracking, verbal processing, and dual task tests used in the AGARD STRES battery. Some comments on measurement strengths and appropriate study designs and methods are included. — Environ Health Perspect 104(Suppl 2):247–273 (1996)

Key words: cognitive tests, psychomotor tests, drug effects, diazepam, performance models, psychometrics, test standardization, study design

Introduction

Psychological performance tests are used to assess the effects of environmental stressors. Most performance tests have found their main application in the study of drug effects, although other stressors, e.g., fatigue, toxic chemicals, are mentioned where appropriate. There is no particular performance emphasis because the tests are intended to have wide applicability. However, vehicle-driving performance is discussed since it has been the subject of a great deal of research, and is probably one of the most important areas of application.

This paper was prepared as background for the Workshop on Risk Assessment Methodology for Neurobehavioral Toxicity convened by the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC) held 12–17 June 1994 in Rochester, New York. Manuscript received 1 February 1995; manuscript accepted 17 December 1995.

Address correspondence to Dr.. Anthony Wetherell, Human Factors Section, Chemical and Biological Defence Establishment, Porton Down, Salisbury SP4 0JQ, United Kingdom. Telephone 1980 613478. Fax 1980 613741.

Abbreviations used: MRT, multiple resource theory; RT, reaction time; PETER, performance evaluation tests for environmental research; STRES Battery, standardized tests for research into environmental stress battery; NF test, number facility test; CTS, criterion task set.

Performance tests differ from most other psychological tests in their design, construction, use, and purpose. These features will be mentioned briefly here and discussed in more detail below. The main difference in design is that other tests, e.g., personality, intelligence, clinical, occupational, are intended to differentiate between individuals, whereas performance tests are intended to differentiate between stressors. To assess individuals, tests should be relatively insensitive to variations in environmental conditions but sensitive to individual differences. To assess environmental stressors, the opposite is true. Generally, tests are suitable for one purpose or the other, but sometimes a test may be found that is suitable for both.

In terms of construction, most tests are usually made to be given only once, or a small number of times, to assess an individual against some known standard or norm. Performance tests are usually made so that they can be administered repeatedly to assess performance over time or with changing stressors. Some performance tests are given repeatedly without change, e.g., reaction time or tracking tests. Other tests contain items that the subject could learn if repeatedly given the same test; e.g., if the same mathematical problems were always given in the same order, then subjects would begin to memorize the answers rather than having to work them out each time. Thus, performance tests have to be constructed to present different problems, but of the same, or at least similar, levels of difficulty, often by randomizing the same problems within a test, or by randomizing items within a problem, e.g., numbers in a mathematical test. This relates to the problems of learning and test reliability.

Sometimes simple randomization is not enough. Baddeley's grammatical reasoning test (*1*) contains problems of different levels of difficulty. If these were simply randomized, it is possible that one version might have a preponderance of easy items early on in the test, while the next version might have a preponderance of difficult items. Performance on the second version would then be reduced, and this might be mistaken for effects of a stressor. The answer is to ensure that all levels of difficulty are equally distributed in each version.

In terms of use, performance tests usually do not have normative databases against which the results can be compared. Thus, reference or control data are usually collected concurrently with test data. For example, performance tests are usually administered to people when they are exposed to the stressor and when they are not exposed or to one group of people who are exposed and another group who are not. For cause and effect to be properly attributed, performance tests are normally used as part of an experimental design that has strict requirements and imposes constraints such as time limits on the test.

Sometimes it is not possible to study a concurrent control group, e.g., in cases where people have been accidentally exposed to a toxic chemical. Here, some form of control group is still necessary, and investigators often employ people in the same job who were not exposed or, if this is not possible, people in similar jobs who were not exposed. This involves a considerable problem in study design.

In terms of application, whereas other tests are designed to assess whether a person meets a certain standard or to diagnose problems, performance tests are designed to assess performance, i.e., whether a person can do a job, e.g., in cases where a diagnosis has been made and treatment has been given, the individual's ability to return to

work needs to be assessed. This relates to the test's validity.

Types of Performance Tests

There are very many performance tests and, while it is not the purpose of this paper to provide a compendium, some examples may help illustrate the variety. More complete reviews of tests used in psychopharmacology are given by Cull and Trimble (*2*), Wittenborn (*3*), and Hindmarch (*4*). Most investigators classify tests broadly in terms of what they appear to measure, e.g., sensory function, central processing, motor function, or perceptual-motor function, but this grouping is often only arbitrary and distinctions are often unclear. With the exception of some tests in which the subject self-generates the stimuli (see below), all tests include some elements of all functions, if only because the subject has to be given some stimulus to work with and has to make some response so that performance can be measured. Thus, a cognitive or higher mental function test also includes perception and response functions, and it is possible, for example, that some tests that claim to measure cognition might actually be better tests of perception.

Examples of the diversity of tests used in psychopharmacological studies include proofreading (*5*), card sorting (*6*), pursuit rotor tracking (*7*), adaptive tracking (*8*), rudder control (*9*), multiple limb coordination (*10*), symbol copying (*11*), absolute auditory threshold (*12*), auditory discrimination (*13*), delayed auditory feedback (*14*), auditory reaction time (*15*), choice reaction time (*16*), ocular convergence (*17*), the duration of after-images (*18*), short- and long-term memory (*19*), verbal learning (*20*), digit span (*21*), muscular grip strength (*22*), body sway (*23*), beam balancing (*24*), digit symbol substitution (*25*), putting caps on ball point pens (*26*), tapping speed (*27*), saccadic eye movements (*28*), the Gibson spiral maze (*29*), galvanic skin response (*30*), electroencephalographic changes (*31*), finding hidden words (*32*), critical flicker fusion (*33*), discrimination conditioning of the eyelid response (*34*), time estimation (*35*), serial subtraction (*36*), the Purdue pegboard test (*37*), concept identification (*38*), digit cancellation (*39*), group vigilance (*40*), category clustering (*41*), spontaneous reversals of the Necker cube (*42*), trigram recognition (*43*), concentration (*44*), logical reasoning (*1*), video games for air combat and slalom driving (*44,45*), navigational plotting (*46*), mental rotation (*47*), Tower-of-Hanoi problem solving (*48*), and creativity (*49*).

It might be asked where all these tests come from: most are purpose designed, although some originate from clinical or experimental psychology. Examples from clinical psychology include the widely used digit span test of short-term memory, which originally formed a subtest of Wechsler's clinical memory scales (*50*). Similarly, the digit symbol substitution test formed part of Wechsler's adult intelligence test (*51*) and has been used by many investigators as a performance test, mainly to measure perceptual skills, but it must be remembered that the test also includes decision making and motor components.

The digit span test has not proved particularly sensitive to stressors; it has been used mainly because it is one of the few short-term memory tests that can be administered repeatedly without task-specific learning. The digit symbol substitution test has proved sensitive to the effects of several drugs, including amylobarbitone (*52*), chlordiazepoxide (*53,54*), diazepam (*52,55,56*), flunitrazepam (*39*), flurazepam (*15,31*), imipramine (*57*), lorazepam (*11*), and nitrazepam (*18,25,58–60*).

An example from experimental psychology is the finding that subjects took longer to read aloud the names of colors written in a different color, e.g., the word "red" written in green (*61*). This effect has been widely used in the study of personality (*62*), perceptual (*63*), cognitive (*64*), and response (*65*) processes, but it has so far been neglected by psychopharmacologists, although it has been used to show an interaction between alcohol and mandrax (*66*) and, with a mirror drawing test, to show that diazepam and nabilone reduced anxiety (*67*).

Use of Performance Tests

Performance tests are used very widely to study the effects of environmental stressors, mostly drugs; the best example is diazepam, which has been perhaps the most widely studied drug for effects on performance. Interest has waned now, but in the late 1960s and 1970s the drug was the subject of perhaps the most intensive psychopharmacological research campaign ever waged.

Effects of Diazepam on Laboratory Performance Tests

Reviews have been published at intervals (*68–73*). The following description is based on the type of performance test used as described by the authors. However, there is some overlap, e.g., the digit symbol substitution test is often viewed as cognitive, but it has a high motor skill component. Also, some tests are conceptually similar but are methodologically different or have been reported differently, e.g., arithmetic can include addition, subtraction, multiplication, or division; tracking can include maze following, pursuit rotors, normal or mirror tracing, and specially designed adaptive tracking tests. In many cases, not only the tests' taxonomies, but also the descriptions of techniques, materials, and procedures leave much to the imagination. Memory tests are particularly difficult to categorize as so many variables are involved, e.g., type of information and degree of meaning, presentation mode and time, retrieval cues, recall modes, and intervening activity and time. Note also that the term psychomotor is often used very loosely to cover not only perceptual–motor skills but also cognitive skills.

In terms of sensory and perceptual skills, perceptual speed is impaired by 10 mg diazepam (*74*); letter searching and cancellation are impaired by diazepam in doses of 5 mg or more (*32,40,75–77*), but can be impaired by doses as low as 2.25 mg (*78*). Lower doses of 2.5 or 5 mg are reported to have no effect (*52*) as might be expected, but a dose of 15 mg has also been reported to have no effect (*79*).

Attention and vigilance are reported to be impaired with doses of diazepam ranging from 2.25 to 10 mg three times daily for 14 days (*52,78,80–84*) and are unaffected in doses ranging from 5 mg to single doses of 20 mg or multiple doses of 5 mg three times daily for 14 days (*40,84–86*).

Tracking is reported to be impaired by doses ranging from 2.25 mg to single doses of 20 mg or multiple doses of 10 mg three times daily for 14 days (*8,75–78,83,84, 86–90*) and to be unaffected by doses from 5 to 20 mg (*84,91–93*). There are even two reports of improvement in tracking with doses of 10 mg (*94*) and 5 mg three times daily for 14 days (*95*).

Symbol copying is not affected by diazepam in doses of 5 and 6 mg (*96,97*), but pegboard and other coordination tests are affected in doses from 10 to 28 mg (*76,79,82,85,98–102*).

Effects on simple reaction time are varied and do not seem to be related to dose (but may be related to different methodologies; see Standardization). Some authors report that diazepam slows reaction time in doses as low as 5 mg (*8,52,83,98,103*), and others say that it has no effect in doses of up to 28 mg (*82,88,96,102,104,105*).

Effects on choice reaction time are similarly varied, perhaps for similar reasons. Impairments have been reported with doses of 10 mg (*74,84–86,89,103*), while no effect has been reported from doses of up to 20 mg (*82,84,92,93*); an improvement has been reported with 5 mg three times daily for 14 days (*95*). Wetherell (unpublished data) used the additive factors method to show that 10 mg diazepam impairs the decision and response stages but not the perceptual stage in reaction time.

Tapping speed seems not to be affected (*52,55,82,93,96,97*) except at doses of 20 mg (*106*).

Arithmetic skills are reported to be impaired by doses from 10 to 28 mg (*30,99,100*) and to be unaffected by single doses of 5 mg to doses of 10 mg three times daily [B Biehl, unpublished data; (*40,79,91,105*)].

Digit symbol substitution is reported to be impaired by doses from 5 to 28 mg (*52,55,102*) and to be unaffected by doses from 2.5 to 20 mg (*79,88*).

Card sorting is impaired by 20 mg (*76*) and unaffected by doses from 10 to 20 mg (*77,80,88*).

Regarding memory, diazepam shows impairments in recall of digits or letters (*40,52,55,106,107*), word lists (*80,102, 106*), addresses and telephone numbers (*76,77,102*), and picture recall and recognition (*108,109*). The effect appears to be not on recall, but on storage processes (*20,106,110*). Other miscellaneous effects of diazepam include impairment of time estimation (*77,98*), but there is no effect on counting or on color naming (*91*).

Effects of Diazepam on Driving Tests

Batteries of laboratory tests are generally reliable, controllable, sensitive, safe, cheap, and convenient. With sufficient imagination, almost any laboratory test can somehow be related to some aspect of driving. However, the predictive validity of laboratory tests is poor, and very few tests have any empirical justification (*72,111–114*). Thus, some investigators have turned to simulators and some to real vehicles.

Using driving simulators, some authors have reported impairments in lane positioning, speed maintenance, and emergency decision making with doses from 5 to 15 mg diazepam (*115*), while others have reported no effect from two doses of 5 mg to single doses of 20 mg (*93,116*).

Using real vehicles, 10 mg diazepam impairs readiness to brake (B Biehl, unpublished data) and the timeliness of decisions to overtake another vehicle (A Wetherell, unpublished data), but ability, confidence, and willingness to drive through narrow gaps can be impaired, improved, or unaffected, perhaps depending on the level of anxiety suffered by the driver (*117*).

After all this work it might be expected that the effects of diazepam on performance are now well known. Unfortunately this is not the case; one reason is that there has been such a variety of tests and results, some contradictory, that all we really know is that diazepam generally impairs performance. However, a recent survey of expert opinions has begun to put some order into the effects of diazepam, and also other drugs, particularly with respect to their effects on driving (*118,119*).

Models of Human Performance

There is no unified model of human performance rather, there are many, although they may be drawn into four main approaches: factor analysis, general information processing, multiple resource/resource strategy, and the processing-stages model.

Factor Analysis

The basis of factor analysis is that a significant correlation between two variables indicates the existence of a common, underlying factor that, at least in part, determines the scores on both variables. Factor analysis begins with a correlation matrix showing the intercorrelations of many variables and attempts to explain these patterns by deriving a smaller number of factors that would generate such a pattern. This is also economical because having to propose a separate factor for every correlation will produce unmanageable numbers with small sets of variables. For example, there are 10 possible intercorrelations with five tests, and in general, $(n \times n-n)/2$ possible correlations with n tests. In factor analysis, the number of factors is generally very much lower than the original number of variables. For example, it may well be found that one single factor could account for most of the variance on perhaps five manual dexterity tests.

The first product of a factor analysis is a factor matrix showing the factor loading of each test on each factor. The factor loading is an estimate of what the correlation would be between the test and a pure test of that factor and indicates the extent to which the test contains, or is loaded on that factor. There is no such thing as a factorially pure test, but some tests can come close.

The next step is to identify each factor. For example, if a factor shows high loadings for tests of basic arithmetic, number series, mathematical reasoning, etc., then that factor might be called mathematical ability. With larger numbers of tests, it is possible that such a factor really contains several component factors such as number fluency, computational ability, etc.

Factor analysts are divided on the philosophical basis of factor analysis. Some contend that factors are real in nature and simply waiting to be discovered. That is, such things as mathematical, verbal, and spatial abilities exist and will be discovered if we are clever enough to use the proper tests. Others argue that factor analysis is merely a way of looking at and summarizing data: factors do not exist in nature but are an interpretation of the data. Interestingly, people in the latter group are usually willing to accept their own results as describing the state of nature, even though the approach is largely atheoretical.

Perhaps the best known exponent of real factors is Guilford (*120*), who proposed a three-dimensional model called the Structure of Intellect, based on his vast experience in the fields of intelligence, creativity, and performance measurement. Guilford's dimensions are *a*) contents, which represent types of information that can be discriminated, e.g., visual, auditory, symbolic, semantic, behavioral; *b*) operations, which are the kinds of intellectual processes that can take place, e.g., evaluation, convergent production, divergent production, memory, cognition; and *c*) products, which are the intellectual outputs of processing, e.g., units, classes, relations, transformations, implications. Thus, Guilford's boxes in his three-dimensional model are real, and he spent much of his research in discovering tests to fill them. For example, Wechsler's digit span test (*121*) could represent an entry in the auditory–memory–units box.

In contrast, the data summary approach starts with only a loose set of hypotheses about the nature of the factors, based on observed consistencies in task performance, and proposes abilities to account for the consistencies. The nature of the ability is then refined by factor analysis. The goal is to select tests such that each underlying factor is represented.

Perhaps the foremost exponent of such task taxonomies is Fleishman, who did a great deal of research to identify major performance factors, generate a taxonomy, and to create a set of rating scales to assess

each element in the taxonomy. Theologus and colleagues (*122,123*) used factor analyses to derive 37 basic human abilities ranging from verbal comprehension to control precision. This number was later expanded to 52 abilities and was published as the manual for ability requirement scales (*124,125*).

Both the real-factor and data-summary approaches have advantages and disadvantages. First, a pure data-summary approach is not feasible because there must be some preconceptions in order to know what tests to include. Second, the data-summary approach depends heavily on the choice of tests; it is possible to omit entire performance areas by failing to include the proper tests. Third, the real-factor approach depends heavily on the investigator's wisdom in including all the relevant dimensions. Also, both approaches suffer from the indeterminacy of factor solutions: the final set of factors depends partly on the method of extracting them and partly on the adjustments that are made in the calculations, called rotations. Guilford (*120*) suggested that, when all else fails, one should "rotate to psychological meaning."

Despite these disadvantages, the factor-analysis approach is useful in that it avoids the subjectivity of trying to predict real-life performance from experimental effects. First, one develops a set of tests to cover the entire spectrum of relevant abilities. Then the tests are related to underlying ability factors by factor analysis—using the coefficients to derive scores on these factors from the test scores. Thus, the effects of stressors on the basic underlying abilities can be determined. If criterion scores from real-life performance can be included, then one can also determine the relevance of experimental effects to real life.

General Information Processing

The literature is increasingly littered with references to information-processing models and tests, as if there were a single, well-defined theory. In fact, there is a large and varied collection of ideas and subtheories, many of which are mutually contradictory. The term information processing seems to be used more as an aid to respectability, or simply because it is fashionable. However, there are two submodels that have been well developed, have substantial empirical backing, and have widespread support among experimental psychologists, although it is also widely accepted that neither provides a complete account of all aspects of human information processing—the multiple resource/resource strategy model and the processing-stage model.

Multiple Resource and Resource Strategy Models

It has long been known that attentional capacity is limited; even the ancient Greeks debated whether it was possible to pay attention to more than one thing at once (*126,127*). Since it is widely accepted that this is not possible, that performance is impaired when more than one task is performed at the same time, a mechanism to allocate the limited attention to competing demands must exist. At first, it was thought that humans had a single, undifferentiated reservoir of resources from which allocations are made to the various tasks, either by intermittent time sharing or by simultaneous apportioning. Thus, when all resources are being used, increased demand from one task can only be met by withdrawing resources from another, causing a decrement in performance of that task. This concept was first proposed by Broadbent (*128*) and was developed during the late 1960s and early 1970s by Moray (*129*), Kahneman (*130*), Norman and Bobrow (*131*), and Navon and Gopher (*132*).

However, the idea of a single pool of attentional resources proved too simple and could not account for some experimental findings; Wickens (*133*) pointed out several. First, some tasks can be time shared with little or no impairment in either (*134–139*). For example, skilled pianists can time share sight-reading music with verbal shadowing with no decrement in either task; skilled typists can similarly time share transcribing written material with verbal shadowing (*134*).

Second, some combinations of tasks show difficulty insensitivity (*140*), in which increasing the difficulty of one task, which should increase its resource allocation, does not interfere with a second task (*141–144*). For example, three different levels of difficulty of a discrete digit-processing task interfered with an additional digit cancellation task, but did not affect simultaneous tracking performance (*142*).

Third, there is a structural alteration effect in which a change in the structure of one of two simultaneous tasks, such as in input or output modality, causes a change in interference between the two tasks, although the difficulty of both tasks remains unchanged (*134,139,140,145–153*).

Fourth, there is an "uncoupling of difficulty" effect where, when two tasks are paired with a third task, the more difficult of the first two tasks actually interferes less with the third task than with the easier one (*138,140*).

These problems may be explained, e.g., by automatization or by assuming additional structures (*130*), but this is unsatisfactory in that every result could be accounted for by assuming a new structure. There are two better alternatives: *a*) assume that there is not one resource pool but several independent ones (the multiple resource model); *b*) assume that processing changes as a result of practice (the resource strategy model).

The multiple resource model (*132,140, 141,150,154,155*) proposes that, instead of a single pool of resources, there are several independent pools. If two tasks draw heavily on the same pool, they will interfere with one another; if the tasks draw on separate pools, they will not mutually interfere. Thus, two tasks will interfere with one another to the extent that they share the same resource pools.

The model postulates that resource pools lie along three dimensions, which were drawn together by Wickens (*140*). The first is processing stages (encoding, central processing, and responding). Several studies have shown that tasks that rely primarily on perceptual processing can efficiently be time-shared with tasks that are primarily loaded on responses (*138*). In contrast, two perceptual or two response-loaded tasks interfere with one another (*135*). Also, difficulty insensitivity is often shown when the two tasks seem to involve different processing stages (*141,143,144*).

The second dimension of resource pools is hemispheres of processing (spatial and verbal). Kinsbourne and Hicks (*154*) showed that interference is greater when a verbal task was combined with a second task in which the right hand (corresponding to the left, verbal hemisphere) was used, compared with the left hand (right, spatial hemisphere). Brooks (*156*) showed that hemispheric specificity could occur with only one task—a task involving spatial working memory could be performed better in combination with a verbal response, and a task involving verbal memory could be performed better with a spatial response. Also, reaction time is slowed when the hemisphere used to process the stimulus is the same as that controlling the response (*151,157,158*).

The third resources pool dimension is modalities of processing (visual and auditory). This is more difficult to establish because the experimental results conflict

somewhat. Some suggest that cross-modulating two tasks is advantageous (*135,145, 146,159,160*) whereas others claim it is not (*161,162*).

A fourth dimension of manual versus vocal responding has been suggested, but has not been fully separated from the spatial versus verbal dimension. An interesting point about multiple resource theory (MRT) is that both encoding and central processing draw on the same resource pool, whereas the processing stage theory considers them as separate stages.

The multiple resource model is not without difficulties; it is limited and it is easy to conceive of resource pools that are not included in the model, e.g., tactile or kinaesthetic modalities. It is possible that some stressors have only minor effects on visual and auditory modalities and might pass unnoticed. The same problem arises when interpreting test results; an effect of a stressor appears to the extent that the tests cover all the resource pools.

MRT is still somewhat controversial. Some people hold that there is a central, undifferentiated resource pool with several independent pools (*163*); others believe that the danger of proliferating pools is too great and that the concept has little or no practical use; yet other people have questioned the notion of completely independent pools. Intuitively, it is difficult to conceive that, for example, divisions of visual attention to spatial and verbal information operate simultaneously but independently of each other. Also, in some cases, performance on two tasks performed together is better than on each task performed separately (*164–167*).

These problems led to some investigators rejecting the notion of a volume of resources in favor of a resource strategy theory, which emphasizes qualitative strategic shifts in performance; performance undergoes fundamental changes as a function of practice (*168*), processing priorities, or information load (*167,169*). In this model, resource is a vague concept referring to almost any processing capability, energetic as well as structural (*170*). Resources are "acquired information about the structure of particular tasks and about the external world which are used by the subject in order to actively control their momentary perceptual selectivity and their choice of responses" (*167*).

Thus, the resource strategy model emphasizes active, top-down control. Furthermore, the locus of control within the human system can vary from time to time during a task depending on the task demands and the systems' idiosyncratic characteristics (*167,171,172*). The strategy model is quite popular, but the framework is almost without predictive power owing to a lack of assumptions. Any result can be made to fit the model simply as another qualitative change (*173*).

Processing Stage Models

The central assumption of the processing-stage model of performance is that a number of mental operations, or processing stages, occur between stimulus and response. A stimulus possesses potential information, and its presentation initiates a sequence of mental operations in which each stage operates on the information it receives and makes the transformed information available to the next stage. These operations take time, which can be measured.

The first experimental studies of information processing stages in reaction time (RT) are attributed to the Dutch ophthalmologist F.C. Donders (*174*) who, citing Johannes Müller's pronouncement of some 25 years earlier that nerve conduction time was infinitely short and could not be measured, pointed out that, by 1850, the German scientist Helmholtz was doing exactly that in frogs and, subsequently, in humans.

Helmholtz found that the muscle contraction in the thumb took longer when the nerves were stimulated at the elbow than when they were stimulated at the wrist; by subtraction, he estimated human nerve conduction velocity at 100 ft/sec. This was surprisingly accurate given the technology of the time, the speed, and the short distance over which it was measured. Helmholtz carried out a similar study of voluntary responses—subjects' moving their hands in response to stimulation of the skin at different distances from the brain.

Donders was also influenced by the work of the French astronomer Hirsch, who found that responses (moving a hand) were slowest to visual stimuli, faster to auditory stimuli, and fastest to tactile stimuli. Donders replicated this work and found that RT took 1/5 second to visual stimuli, 1/6 second to auditory stimuli, and 1/7 second to tactile stimuli. However, neither Donders nor Hirsch took stimulus intensity into account. Now it is well known that RTs decrease as stimulus intensity increases, partly because nerve conduction speed across synapses increases with stimulus intensity.

Given Helmholtz's measures of nerve conduction speed and Hirsch and Donders' absolute RTs, Donders reasoned that nerve conduction time could only account for a small part of the total RT and devised several experimental paradigms to measure the duration of the mental processes involved. In one set of paradigms, Donders stimulated subjects' feet and asked them to respond by moving the hand on the same side, showed red or white lights and asked subjects to respond by moving their hands, or presented two letters of the alphabet and asked subjects to pronounce the name of the letter. RT was measured under two conditions. In the first, the subjects knew which stimulus was to be presented (simple RT); in the second, they did not (choice RT). Thus, in choice RT, the subjects had to perform two additional mental operations: identify which stimulus had been presented and select the appropriate response. Donders found that choice RT for foot stimulation was longer than simple RT by 67 msec; he attributed this to "the time required for deciding which side had been stimulated and for establishing the action of the will on the right or left side." Choice RT with red and white lights was longer than simple RT by 154 msec and, with letters, by 166 msec.

Donders devised another series of experiments to isolate stimulus recognition and response selection stages. To measure recognition time, he measured RT under two conditions: first, subjects had to push a button when they saw a light (no recognition); second, the light could be one of two colors and subjects had to respond only to one (recognition needed). To measure response selection, he again measured RT under two conditions: first, subjects were presented with either of two signals but had to respond only to one (no response selection); second, subjects had to respond differentially to both stimuli (response selection needed).

Thus, by subtraction, Donders hoped to measure the duration of these processing stages. His idea was good but wrong. He did not realize that stages cannot be inserted or removed without changing the nature of the task. For example, there is a sense in which subjects can be said always to have to select a response—whether to respond or not—even in simple RT. Thus, it cannot be said that a response selection stage was present in one condition but not in another. At the turn of the century, Donders' subtractive method was further criticized on the basis of introspective reports that stages cannot be added without

affecting the time to complete other stages. Donders' method was then forgotten until the 1960s, when it was revived by Sternberg (*175–179*).

Sternberg (175–179) accepted that stages could not be inserted or removed without affecting the nature of the task; instead, he proposed that the amount or difficulty of processing at each stage could be manipulated by using an additive factors method. Sternberg's studies were largely concerned with memory search, but additive factor principles can be applied to any task.

First, the experimenter decides what stages might be present in a task and then chooses independent variables that are expected to affect particular stages. If the variables have additive effects, then they are inferred to affect separate stages; if the variables interact, then they are inferred to affect at least one common stage. For example, stimulus intensity might be expected to affect detection, whereas response compatibility (e.g., dominant or nondominant hand) might be expected to affect response selection in an RT task. Logically, the difficulty of the response should not affect the ability to detect a stimulus, and the ease of detecting a stimulus should not affect the response.

If this is true, then plotting RT as a function of stimulus intensity should give parallel curves. If it is untrue—if stimulus intensity affects not only the time to detect a stimulus, but also the time to select a response, e.g., if it speeds up both—then the two variables will interact and the two curves will not be parallel.

Thus, the loci of effect of stressors may be found by searching for interactions between the stressor and variables chosen to affect particular stages. Of course, it should first be determined that the chosen variables do not themselves interact. For example, Wetherell (unpublished data) showed that atropine impaired the perception and decision stages in reaction-time performance, whereas diazepam impaired the decision and response stages.

It is important to remember that "the additive factors method cannot distinguish processes but only processing stages" (*178*). As with all models, the additive factor model has been criticized. There are several research findings, e.g., serial position effects, that cannot be reconciled with the theoretical basis of additive factors (*180,181*), and the performance profile could be due to other causes, e.g., stimulus range effects (*182*). It is logically possible for two variables to affect the same stage and yet show additive effects; interactions could be masked if the two variables affect the stage in opposite ways, and two or more processes could proceed in parallel, e.g., response selection and stimulus identification. Also, patterns of interactions are not always sufficient for estimating the number of stages, and uncritical use of the model could give misleading conclusions about the loci of effects (*183*). However, there can be no argument that if several factors independently affect performance on a task and a drug interacts with one of them and not the others, the drug affects that aspect of task performance and not the others, whatever that aspect might be called and whatever the task's theoretical basis.

The additive factor and resource strategy models are not compatible (*167,172*) because the latter holds that changes in task demands alter the architecture of processing sequences. However, the additive factor model is compatible with the resource volume model (*173*) since the energetical supply to the stages can be considered as resources and attentional aspects can be incorporated. Thus, it should be possible to discover which stages are affected by energetic variables, and the effects of cognitive states such as knowledge of results and time pressure could be studied. The additive factor model, therefore, may be useful in identifying not only processing stages, but also the effects of resource allocation. However, Gopher and Sanders (*173*) argue against combining the two models because they have different methodologies and applications.

Psychometrics

Any psychological test should be valid, reliable, and sensitive, i.e., it should measure what it purports to measure, do so consistently, and be capable of detecting changes in what it measures. These principles are commonly applied in many areas of psychology such as personality, intelligence, and clinical and occupational testing but are rarely applied to performance assessment and hardly at all to the assessment of drugs on performance (*3,4,184*).

Parrott (*185*) analyzed a sample of 38 papers from 16 journals covering 115 tests and all the major drug types over the period from 1972 to 1988. He took only one study from each laboratory, since any given research group tends to use the same set of tests. He found that none of the papers documented either reliability or validity, although nine mentioned that validity had not been established, and one mentioned that reliability should be investigated. Very few tests were used by more than one laboratory, although this was difficult to assess since test descriptions were usually very brief.

The advent of personal computers is at least partly to blame. It is true that computers have improved measurement accuracy and formalization of procedures, but their very versatility seems to have seduced psychologists into producing an ever increasing number of performance tests, many of which appear to be more demonstrations of elegant programming and beautiful graphics than serious attempts to measure performance. It is almost a case of if it can be done, it has been, or will be, done. "If a thing exists it exists in some amount. If it exists in some amount it can be measured" (*186*).

Discussions of psychometrics have been given by Guilford (*120*), Cronbach (*186*), Kelly (*187*), Anastasi (*188*), Kline (*189*), Jones and Appelbaum (*190*) and, with particular reference to psychopharmacology, Parrot (*185,191,192*); only brief comments on reliability, validity, learning, and test standardization are necessary here. First, however, a brief description of measurement is appropriate.

Measurement

The basis of psychological measurement is the assigning of numerical values to behavioral events such that differences in behavior or performance are represented by differences in test scores. This implies an underlying logic, or set of rules, that governs the relationship between numbers and events, and the kinds of statistical treatment of test scores that are permissible depend on this logic. Detailed accounts of measurement scales are given by Stevens (*193*) and Siegel (*194*), but some comments are appropriate here.

Broadly, there are four different ways of quantifying a variable, or four different types of scale or measurement strengths, that can be used. The first is simple categorizing of subjects or responses into classes, e.g., hits and misses, passes or fails. This is known as nominal measurement, and the data are not in the form of scores but in the form of frequencies in different classes. An example of nominal measurement is classifying subjects as neurotic or normal, extraverted or introverted.

The second form of measurement, stronger than categorization, is ordinal, in

which the subjects or responses are placed in rank order with respect to the variable concerned. For example, shots at a target can be ranked according to their distance from it, or subjects can be ranked according to the number of test items they answer correctly. Runners in a race are ranked according to their finishing order. Thus, ordinal measurement can show whether one score or subject is better or greater than another, but it does not say by how much because the magnitude of the differences between scores is not constant. Intelligence test scores are often ordinal: an individual with an IQ of 150 is not necessarily twice as bright as one with an IQ of 75; the difference between an IQ of 90 and one of 100 is not necessarily the same as the difference between an IQ of 100 and one of 110.

The third and next stronger form of measurement is interval in which the differences between measurement units are constant, e.g., temperature is measured on an interval scale. The fourth and strongest form of measurement is ratio, which is an interval scale with a true zero point, e.g., length, response time (even though it is not possible for a subject to respond in zero time, the scale does have a zero point).

The different measurement scales require different mathematical and statistical treatments. Briefly, arithmetic means, standard deviations, and parametric tests of statistical inference such as Student's *t*-test and analysis of variance require ratio or at least interval measurement. They also require that the data be continuous and normally distributed. Ordinal and nominal data require medians, modes, ranges, and nonparametric statistical tests (*194*). Nonparametric tests can also be used with interval or ratio measurements, but they are generally less powerful.

Reliability

Given that the purpose of any psychological test is to discriminate between subjects or scores, it is most important that the test does so consistently, i.e., that it is reliable. When a group of subjects takes a test, the resulting distribution of scores will be a function of stable differences between them or between stressors and other differences owing to a wide range of other factors, collectively called error variance.

Thorndike (*195*) listed several possible sources of variation in performance on a test:

- Lasting and general characteristics of the individual
  - General skills (e.g., reading)
  - General ability to comprehend instructions, testwiseness, techniques of taking tests
  - Ability to solve problems of the general type presented in the test
  - Attitudes, emotional reactions, or habits generally operating in situations like the test situation (e.g., self-confidence)
- Lasting and specific characteristics of the individual
  - Knowledge and skills required by particular problems in the test
  - Attitudes, emotional reactions, or habits related to particular test stimuli
- Temporary and general characteristics of the individual (systematically affecting performance on various tests at a particular time)
  - Health, fatigue, and emotional strain
  - Motivation and rapport with experimenter
  - Effects of heat, light, ventilation, etc.
  - Level of practice on skills required by tests of this type
  - Present attitudes, emotional reactions, or strength of habits.
- Temporary and specific characteristics of the individual
  - Changes in fatigue or motivation in this test (e.g., discouragement resulting from failure on a particular problem
  - Fluctuations in attention, coordination or standards of judgment
  - Fluctuations in memory for particular facts
  - Level of practice on skills or knowledge required by this particular test (e.g., effects of special coaching)
  - Temporary emotional states, strengths of habits, etc.
  - Luck.

Reliability can be measured in several ways: *a*) test–retest reliability (or coefficient of stability) measures variation with time or test sessions; *b*) internal reliability measures content variance (variation between items in the test), often as split-half reliability or coefficient of equivalence; *c*) alternate form reliability measures further aspects of content variance and also imperfect matching of different forms of the test; *d*) interrater reliability measures variation in scoring techniques; and *e*) interpresenter reliability measures variation in ways of presenting the test.

Reliability is important not only for the stability of tests, but also for repeated-measure designs in which stable test–retest correlation and stable variance are technical requirements of analysis of variance. All measures of reliability are affected by other sources of error variance, including subject motivation, distraction, fatigue, and learning.

One might think that high reliability is a disadvantage in tests designed to detect changes in a subject's performance with and without stressors. However, this would confuse reliability and sensitivity. Ideally, tests should have high test–retest reliability, indicating that they are stable under constant conditions, together with high internal reliability, but they should reflect the effects of stressors to which it is intended to be sensitive.

By way of illustration, Bittner et al. (*45*) collected test–retest reliability data on 140 performance tests, and some examples are given in Table 1.

Validity

A test is valid if it measures what it claims to measure: "For so it is oh Lord my God, I measure it, but what it is that I measure I do not know" [St Augustine]. This sounds obvious, but it is not always easy to achieve. Validity has to do not only with the test itself, but also with how it is used: a test may be valid for one purpose and invalid for another. Validity is a complex subject; it is normally classified into three main types: content, criterion (concurrent and predictive), and construct validity. A fourth type, face validity, is also usually

Table 1. Test–retest reliability coefficients normalized for a 3-min period.[a]

Test	Reliability coefficient
Stroop: time to name color words	+0.97
Logical reasoning time	+0.93
Arithmetic vertical addition	+0.90
Letter search	+0.87
Aiming: hand–eye coordination	+0.87
Code substitution	+0.84
Perceptual speed: number comparison	+0.84
Four-choice reaction time	+0.80
Sternberg item recognition (set 4)	+0.80
Manikin test: mental rotation	+0.79
Minnesota manual dexterity: turning	+0.64
Air combat maneuvering: Atari simulation	+0.63
Letter classification memory: LTM name retrieval	+0.55
Target tracking accuracy: two-dimensional	+0.52
Memory: free recall	+0.52
Stroop: color/word naming difference	+0.47
Choice reaction time: information slope	+0.41
Navigational plotting accuracy	+0.40
Sternberg item recognition: information slope	+0.11

[a]Data from Bittner et al. (*45*).

considered, although strictly it is only concerned with what a test appears to measure and not with what it does actually measure.

Content Validity

Content validity reflects the extent to which a test adequately covers a particular area of interest. It is straightforward in narrow, well-defined areas, e.g., weapon aiming, missile tracking, but becomes more difficult as the areas become broader, e.g., general military skills. Some investigators have chosen to use only one test whereas most have used batteries of tests, but in both cases the problem lies in deciding which psychological functions to include.

With regard to batteries of tests, Wesnes et al. (*196*) suggested four psychological functions that should be covered: attention, selecting information from the environment; cognition, processing this information; memory, storing the information; and response, physical coordination in responding to the information. Hockey and Hamilton (*197*) suggested five functions: alertness; selectivity; speed; accuracy; and short-term memory. Cull and Trimble (*184*) also proposed five: attention and sensory processing; mental speed; central cognitive processing; memory; and perceptual-motor performance. Parrott (*198*) proposed six psychological functions: sensory reception and attention; arousal and alertness; simple information processing; complex information processing and cognition; memory storage; and simple psychomotor performance.

Holding (*199*) proposed four perceptual-motor groupings: simple perceptual, e.g., vigilance; simple motor, e.g., tapping; complex perceptual, e.g., air traffic control; and complex motor, e.g., aircraft piloting, and suggested that verbal-intellectual skills required further tests. Other taxonomies have been proposed (*2,4,124*), but, while they agree generally, they disagree in some particulars. For example, Hindmarch (*4*) listed digit symbol substitution as a sensory task whereas Parrott (*198*) used it as an information-processing task; Cull and Trimble (*2*) called it a psychomotor task.

Bittner et al. (*45*) carried out one of the most exhaustive psychometric assessments of performance tests for use in their performance evaluation tests for environmental research (PETER) battery. They selected 45 tests from a literature survey of 145 tests, assessed the tests' reliability (see Table 1), used factor analysis to identify similar test subsets, and used reliability–efficiency data to select the most sensitive test versions. The PETER battery consisted of five tests: logical reasoning (left hemisphere cognitive); pattern comparison (right hemisphere cognitive); code substitution (memory/perceptual); aiming (fine sensory-motor control); and spoke control (gross psychomotor). However, despite the psychometric care taken in its design, the PETER battery has several omissions (e.g., attention and vigilance, memory, arousal, simple and complex psychomotor skill), and sensitivity to stressors such as fatigue and drugs was not considered.

With regard to using only one test, Wood et al. (*200*) used a tracking task to assess motion sickness preventatives for astronauts, found no effect, and concluded that "these drugs should produce no significant performance decrement in an operational situation." This is clearly inadequate. but sometimes using only one test can be justified. For example, Hindmarch and Clyde (*42*) suggested that discrete choice reaction time could provide an index of overall sensory-motor integrity. Hockey and Hamilton (*197*) described several studies that used the Wilkinson continuous choice reaction time test, and suggested that different stressors produced different patterns of effects. The additive factors paradigm offers perhaps the best rationale for using a single test since it has been shown to differentiate the loci of action of stressors.

Criterion Validity

Criterion validity reflects the extent to which the test correlates with some criterion of real-life performance. Criterion validity is called concurrent validity if both test and criterion performance are measured at the same time and predictive validity if criterion performance is measured after test performance, i.e., the extent to which the test predicts real-life performance.

The measurement of criterion validity sounds easy but entails several problems. First, a suitable real-life activity has to be selected and justified to the sponsor (the infantry might not accept the investigator's choice of artillery operations).

Second, a reliable and objective performance criterion must be chosen. Sometimes this is easy: artillery operations consist of objectively measurable skills such as siting, loading, and firing the weapon, defined performance metrics (usually speed and accuracy), and objective performance criteria (to hit the target with the correct munition at the correct time).

Sometimes choosing reliable and objective performance criteria is difficult—helicopter piloting is objectively measurable—but which criterion should be chosen—the ability to maintain steady flight, or hover, correcting for wind gusts? Flying a helicopter is rarely an end unto itself, rather, it is a means to an end. If the end is to arrive at a destination, then map reading and route finding are important as is the ability to land the machine safely, e.g., using a ground-controlled radar approach. If the end is to observe enemy activity (as in artillery spotting) or to hit a target, then the pilot is only a chauffeur, and the observer's vigilance or the missile operator's tracking are the criteria.

Sometimes choosing performance criteria is almost impossible. A brigade commander has a very demanding job; mistakes can kill not only him, but also thousands of others. The military is, therefore, very interested in his performance. Unfortunately, there is no defined performance metric—most brigade commanders appear quite inactive most of the time, but this is not to say that they are doing nothing. The only performance criterion is whether he wins or loses the battle, but even this is not easy to tell: is it in terms of lives or equipment lost, or objectives gained despite losses? What if he only partially gains his objective?

Despite these problems, several validations of laboratory or simulator tests in terms of real-life performance have been made. For example, Henry et al. (*201*) validated multidimensional pursuit and complex coordination tasks against performance on an aircraft simulator. Billings et al. (*202*) found that secobarbital impaired the tracking accuracy and airspeed control of pilots in light aircrafts and in simulators and had found earlier that the same flying tasks were impaired by alcohol (*203*). Seashore and Ivy (*204*) found that stimulant drugs (to combat sleepiness) improved performance on a range of laboratory tests, including choice reaction time, critical flicker fusion, and target tracking, and on real military tasks, including tank driving, truck driving at night, and prolonged guard duty. However, Seashore and Ivy (*204*) found several differences between their laboratory and real tasks; and Billings et al. (*202*) also found several differences between the simulator and the real aircraft and warned against direct extrapolation.

The most widely studied criterion validity problem involves driving. Hansteen et al. (*205*) found that cannabis and alcohol

impaired laboratory target tracking and subjectively assessed steering accuracy in a real car. De Gier et al. (*206*) found that diazepam impaired performance on a laboratory attention task and on subjectively assessed real-life driving.

Linnoila and co-workers (*74,86,89,90, 95,207–209*) are perhaps the foremost proponents of this method. Their concurrent pursuit tracking and choice–reaction time tasks are based on previous evidence (*210*) that these measures are correlated with the accident records of 100 Helsinki bus and tram drivers. However, the validity of Linnoila's methods has been questioned, since they apply only to a parochial sample (*114*) driving under time pressure and often adverse conditions (*113*).

Although these finding seem to indicate that laboratory tasks do have criterion validity, all that has actually been found is that some drugs affect some aspects of both laboratory and real-life performance. In some cases these are obviously related, e.g., laboratory tracking and steering accuracy, although most authors did not report any correlations. In other cases the relationship appeared only coincidental, e.g., false alarms on an attention test (*206*), and other aspects of the tasks were affected differently. Generally, laboratory tests are considered poor predictors of car driving ability (*72,112–114*).

Thus, it is difficult to assess the criterion validity of a laboratory test in a single study, but it may be possible to do so over several studies by meta-analysis. For example, several studies have shown that diazepam impairs driving ability in real cars. O'Hanlon et al. (*211*) found that it impaired lateral positioning, de Gier et al. (*206*) found that it had similar effects assessed subjectively, and Wetherell (*117*) found that it affected confidence, skill, and willingness to act when judging whether it was possible to drive through narrow gaps. Similarly, several authors have found that diazepam impairs laboratory tasks. In contrast, clobazam generally appears to have no effect on either laboratory tasks (*36,212*) or real driving (*213,214*).

However, finding that a laboratory test and a measure (however obtained) of real-life performance are both affected by a drug does not mean that the laboratory test has criterion validity. The driving tests used might be able to measure, for example, the tranquilizing effects of diazepam, but that does not validate the tests as measures of driving or as measures of tranquilization; it simply validates them as a measure of the presence of a drug. Sensitivity to a stressor is a necessary condition, but it is not sufficient by itself.

Construct Validity

Construct validity refers to how well the test reflects an accepted model and fits with other supporting evidence. Theories and models of human performance have already been discussed, but some further comments are appropriate here, in line with the examples from psychopharmacology used to illustrate other aspects of validity. Psychopharmacologists tend to base their methods on taxonomies or two groups of models.

Regarding taxonomies, several have been proposed: mental versus physical, cognitive versus noncognitive, perceptual versus motor, simple versus complex, skilled versus habitual, and open versus closed (*199*). Some are listed above (Content Validity), and Parrott (*192*) proposed another taxonomy, perhaps more appropriate for psychopharmacologists: *a*) stimulus reception (acuity, sensation, perception); *b*) attention [detection of targets in a matrix of rapidly presented repetitive stimuli, e.g., letter cancellation; rapid information processing (*49*)]; *c*) vigilance [detection of uncertain and infrequent stimuli over a prolonged period, e.g., auditory vigilance; Mackworth clock test (*215*)]; *d*) simple information processing (code substitution, symbol coding, arithmetic, Stroop name identification and color identification); *e*) complex information processing (thinking and, logical reasoning, mental rotation, concept identification, creativity, judgement); *f*) cognitive attention, distinguishing cognitively confusing stimuli, e.g., Stroop color-word difference (*61*); *g*) memory [digit span, recognition, recall, consolidation, retrieval, Sternberg test (*216*)]; *h*) simple psychomotor skill [tapping, aiming, simple reaction time, choice reaction time, continuous reaction time (*217*), unidimensional target tracking, steadiness, balance]; *i*) complex psychomotor skill (manual dexterity, two dimensional coordination, multidimensional tracking, complex choice reaction time, piloting, complex simulator performance); and *j*) psychophysical, physiological and subjective measures [often included in performance test batteries (*218*) to indicate alertness and arousal]—critical flicker fusion and saccade velocity, heart rate, evoked cortical potentials (often used to assess workload), feeling state questionnaires.

Regarding models, the two main groups are those derived from Donders (*174*) and Sternberg (*177*) and those derived from Broadbent (*128*). Donders' and Sternberg's models are discussed in detail below; the Broadbent model was originally conceived to explain selective attention effects but has been developed by Broadbent himself (*219,220*) and by several psychopharmacologists for their own purposes [e.g., (*4,196,209,221*)].

Face Validity

A test has face validity if it looks like what it is supposed to measure (e.g., a tracking test, especially if controlled by a steering wheel, is a face valid measure of car driving ability); if the tracking is in two dimensions and is controlled by a joystick, then it could be a face valid measure of flying an aircraft. Mackworth's clock test of vigilance (*215*) was designed to measure radar scanning ability; it looked like the real task and was, therefore, face valid.

Face validity is the least important validity test in one sense, but perhaps the most important in another. It is least important because it is only a trivial aspect of the test (*189*); it is not even a validity in the technical sense because it refers not to what the test actually measures, but to what it appears, superficially, to measure (*188*). It is perhaps the most important because subjects, sponsors, and fundholders can accept and relate to it better.

Laboratory tests, discussed above, can be fairly easily constructed with face validity (e.g., tracking tests can be made to look like driving or guided missile control tasks; vigilance tests can be made to look like production-line quality control inspection tasks; aiming tests can be made to look like weapon-aiming tasks. However, there are problems: face validity is sometimes only in the eye of the beholder and it does not guarantee criterion validity.

For example, the military considers marksmanship to be of great importance, and it is easy to construct a test that looks good to a military sponsor. But lining up symbols on a computer screen does not mean that the subject will perform the same when required to aim and fire his weapon in the noise and danger of battle, or even in practice on a shooting range. There is also a deeper problem. Marksmanship might have been important long ago when muskets were cumbersome, inaccurate, and slow to reload, but it is less important nowadays with automatic, rapid-firing small arms. It could be argued that a

personal weapon is really to improve the owner's morale and confidence, and when used in anger, for distracting the enemy or keeping his head down. If you really want to kill the enemy, then you call up something more efficient like artillery or aircraft. Thus, marksmanship might not be the issue: speed of reloading, or disassembling, cleaning and reassembling the weapon are more important.

Obviously, the more a test looks like the real thing, the more face validity it has. Laboratory tracking tests can be controlled by steering wheels and foot pedals, and their displays can portray roads and other vehicles. More sophisticated details can be added until what was a laboratory test might be better called a simulator. Simulators have a great deal of face validity; they are often used for training, but some, particularly driving simulators, have been widely used to measure performance.

Several drugs have been shown to impair performance on driving simulators, including alcohol either alone (*222–225*) or combined with other drugs such as cannabis (*226,227*); amitryptiline (*228,229*); cyclizine (*230*); chlorpromazine and thioridazine (*231*); β-adrenergic blockers (*232*); meclastine, cyproheptadine, and pheniramine (Landauer and Milner, unpublished data); diazepam and haloperidol (*93*); and diazepam (*92*). Other drugs include cannabis alone (*233*); anesthetics (*234,235*); medazepam (*236*); meprobamate (*237*); meprobamate, phenobarbitone and chlordiazepoxide (*238*); diazepam and secobarbital (*115*); and cigarette smoking (*239*). Some idea of what these authors thought of the criterion validity of their simulators may be seen from the titles of their papers: some say "effects of...on driving", but most say "effects of...on a skill resembling driving", or "on a driving simulator".

Klonoff (*240*) found that cannabis impaired driving both on a closed course and in the streets of Vancouver; while driving a real vehicle has even more face validity than a simulator, it is still no guarantee of criterion validity. Most-real vehicle studies have been carried out on a private, closed course or on the highway. Closed-course driving permits some degree of experimental control but is artificial. Highway driving is largely uncontrollable and is often artificial since the subject is aware that he is being studied, either by instrumentation or by human observers. In both cases, the criterion measures, i.e., those that are important are still open to argument.

Several drugs have been shown to affect closed-course, gymkhana-type driving tests. Maneuvering tests are perhaps the most widely used and can be impaired by phenobarbitone, but not by chlordiazepoxide unless it was combined with alcohol (*241,242*), amylobarbitone, trifluoperazine, and haloperidol (*243*), lorazepam (*214*), but not clobazam (*214,244*), imipramine but not viloxazine (*245*) diazepam alone (Wetherell, unpublished data) and with and without alcohol (*246*), atropine (*247*), and the anticholinesterase sarin (*248*).

Gap judging can be impaired by alcohol (*249*), alcohol and chlordiazepoxide (*243*), imipramine (*245*), sarin (*248*), diazepam (*117*), and a nonpharmacological example, concurrent performance of Baddeley's (*1*) verbal reasoning task (*250*).

Car-following situations account for a high proportion of total vehicle involvement in road traffic accidents [Sabey, unpublished data; (*251*)] and have been studied in terms of highway traffic flow (*252*), perceptual cues used by the following driver (Jansseu, unpublished data), and elected headways (following distances) in experimental situations (*253*) and on the highway (*254*). However, very few on-the-road studies of this important aspect of driver behavior have been reported. Alcohol increases mean headway in daytime but not at night (*224*), but atropine (*247*) and diazepam (Wetherell, unpublished data) both decrease mean headway, especially when the driver is preparing to overtake the other car.

Route finding and following have been studied very little, but they do represent the main reason why vehicles are used—to get from one place to another. Atropine impairs drivers' ability to navigate from memory of verbal directions (*247*); diazepam impairs abilities to navigate from memory of both verbal and graphic directions (A Wetherell, unpublished data). Both drugs seem to affect storage rather than retrieval.

Studying drug effects with real cars on real highways can be fraught with safety and legal problems, but some investigators have successfully managed it. Perhaps the foremost proponent of this method is De Gier et al. (*206,255*), who used experienced driving instructors to rate subjects' performances.

Standardization

Most psychological tests (e.g., personality, intelligence, clinical, occupational) are routinely given in standardized versions, and users would not dream of altering them for fear that it would affect the tests' reliability and validity. Performance tests may appear to be standardized, since the same names keep appearing in the literature (e.g., choice reaction time, mathematical processing, logical reasoning), but many of these tests are the same in name only. Even a simple reaction-time test can vary from laboratory to laboratory in several ways, for example, the size, contrast, color, and complexity of the stimulus; the font of any alphanumeric characters; whether the subject is warned; the interval between the warning and the stimulus (called the foreperiod); the time allowed for a response before the next stimulus is presented; the type of response required (e.g., vocal, pressing or releasing a key); the pressure required to operate the key; the amount of travel; where the subject rests his fingers. Similar, or even greater, variation can be seen in other tests, and most tests are so poorly documented that it is impossible to discover exactly what the investigators did.

The reasons are at least 3-fold. One reason is that there is no unified theory or model of human performance. Another reason is perhaps that there has been a long and jealously preserved tradition for each research group to devise its own tests, which have proved useful for its own purposes. The third reason tests are poorly documented is the advent of personal computers, which can seduce the psychologist or programmer into adding or changing features to make the program more efficient, or simply to make the test look better or more impressive. Before computers, many tests were given in paper and pencil versions, which probably forced more standardization, but even then, some investigators would retype, reformat, abbreviate, or expand stimulus sheets. One major result of this lack of standardization is the disagreement and sometimes contradiction of results between research groups.

Some tests have tended to be adopted and used in at least reasonably well-standardized versions. Examples include the continuous, four-choice reaction time test of Wilkinson and Houghton (*256*), itself derived from Leonard's (*217*) five-choice reaction time test; the rapid visual information processing test (*257,258*), derived from an earlier test of vigilance; the "Leeds Psychomotor Tester" (*42*); and Baddeley's grammatical reasoning test (*1*), which tends to be used in its original format, although perhaps the most extreme distortion of the original has been chosen for inclusion in perhaps

the most serious attempt at standardization (see below).

One notable and successful attempt at standardization is the STRES (standardized tests for research into environmental stress) battery (*259–261*). This was the product of the NATO Advisory Group on Aerospace Research and Development Working Group 12, which spent 2 years standardizing seven commonly used performance tests.

Working Group 12 (*259,260*) did not design new tests but selected existing ones based on their track record. The criteria for selection were evidence of reliability, validity, and sensitivity; a documented history of application to the assessment of a range of stressor effects; short duration (maximum 3 min); language independence; sound basis in human performance theory; and the ability to be implemented on easily available computer systems.

The STRES battery comprises six individual tests (reaction time, memory search, mathematical processing, spatial processing, grammatical reasoning, and unstable tracking) and one dual-task test (a combination of memory search and unstable tracking). Two of the tests (reaction time and memory search) are based on the information-processing model and use additive factor methods to differentiate processing stages; the other five tests are more traditional tests of higher mental function or psychomotor skill, identified by factor analyses or based on resource theory.

The standardization included specification of the software, stimulus display, response devices, testing environment, training requirements, subject instructions, and data collection and exchange formats. Examples of the specifications for stimulus display elements were that they should be white on a dark background with a ratio of stimulus to background of between 7:1 and 12:1. Alphanumeric characters should subtend a vertical visual angle of 15 to 20 minutes of arc at a recommended viewing distance of 0.6 m. Examples of specifications for the response keys were that they should be nonlatching, push-to-make switches with a travel of 3 mm and an actuating force of 0.30 to 0.35 N. Examples of the joystick specifications were that it should have a 30° range of movement left and right, the friction should not exceed 50 g and it should be constant over the range of travel, and analogue-to-digital conversion should be at no less than 8-bit resolution. Full specifications are given in AGARDograph 308 (*259,260*).

Working Group 12 (*259,260*) arranged for two organizations to coordinate and manage databases: Laboratoire d'Anthropologie Appliquee (LAA) in Paris, for Europe and the Crew System Ergonomics Information Analysis Center (CSERIAC) at Wright-Patterson Air Force Base, Ohio, for North America. These organizations also act as clearing houses for information on the STRES battery (*260*).

The STRES battery is now being used increasingly to study a wide range of stressors, and the tests are described below as an illustration of the design, construction, use, and purpose of performance tests [full specifications are given in AGARDograph 308 (*259,260*)].

Reaction Time

The early background to RT tests is given above. Since Donders' work, RT has become perhaps the most widely used test of performance, albeit in a wide variety of forms. RT tests come in two types: simple, with one stimulus and one response, and choice, with more than one stimulus and response.

The test used in the STRES battery is based on the Dutch TNO Taskomat Battery (*262*) and is probably the most highly developed RT test available. It includes both simple and choice types and also uses additive-factors methodology to differentiate the information-processing stages involved. Five stages have been identified, and four variables have been devised to affect them (Table 2).

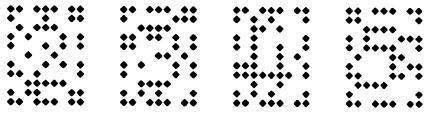
The stimuli consist of the digits 2 to 5 formed from small diamonds and enclosed in a rectangular frame also formed from diamonds.

Table 2. Processing stages, with variables to affect them in the STRES RT Test.

Stage	Variable
Stimulus processing or encoding	Stimulus quality
Response selection	Stimulus–response compatibility
Motor programming	Response complexity
Motor activation	Time uncertainty
Response execution	Response complexity

These stimuli can be presented either on the left or on the right of the monitor screen, and subjects respond by pressing one of four keys, using the index and second fingers of both hands—second finger of left hand for digits 2 and 3 appearing on the left, index finger of left hand for digits 4 and 5 appearing on the left, index finger of right hand for digits 2 and 3 appearing on the right, and second finger of right hand for digits 4 and 5 appearing on the right.

Encoding can be affected by stimulus quality, which can be degraded by moving 10 diamonds from the frame towards the digit, for example,



This preserves figure-ground contrast and ratio.

Response choice is affected by stimulus–response compatibility, which can be made more difficult by swapping hands: left hand for stimuli appearing on the right and right hand for stimuli appearing on the left. Motor programming (and response execution) is affected by response complexity. Instead of pressing a key once, the subject has to press three keys: first the normal key, then the other key for the same hand, and then the normal key again. Motor activation is affected by time uncertainty: stimuli are presented at irregular intervals.

Split-half reliabilities for the test range from 0.81 to 0.92, depending on the block. Split-half reliabilities for the difference scores corresponding to the particular processing stages range from 0.62 to 0.74. The split-half reliability for response execution time was 0.94 (*259*). The validity of the test depends upon the validity of the additive factor method, the rationale of which is that two variables are inferred to affect separate processing stages if they have additive main effects and to affect at least one common stage if they have interactive effects. Thus, to identify the loci of action of an unknown variable such as a drug, the task variables should not themselves interact, and this has been shown to be true (*259*).

The test has shown nonspecific sensitivity to a variety of stressors such as fatigue, sleep loss, aging, brain damage, vitamin deficiency, and a range of drugs including barbiturates, amphetamines, and antihistamines (*263–268*). Regarding specific effects on particular stages, stimulus encoding is reported to be affected by brain concussion

(*269*), sleep deprivation (*270,271*), and barbiturates (*272,273*). Response choice is affected by brain concussion (*269*) and sleep deprivation (*270*). Motor activation is affected by brain concussion (*269*) and sleep loss (*274*).

Mathematical Processing

Mathematical processing tests, unless they are for very specialized use, are actually arithmetical tests in which the subject has to add or subtract, or more rarely, divide or multiply numbers. One of the first mathematical processing tests to receive proper psychometric attention and documentation was the paper and pencil number facility (NF) test (*275*), which required subjects to add three one- or two-digit numbers and write the answers in boxes. The test was standardized on U.S. servicemen (*276–279*), and some work was done in the United Kingdom (*280*).

The NF test showed an interesting problem when attempts were made to computerize it (*281*). The paper and pencil version allowed subjects to use any addition strategy, but the computer program constrained subjects to add the numbers in what was thought to be the most common way—least significant digits first. However, there was a small but significant number of people who had learned the efficient strategy of estimating using most significant digits first and then adjusting by adding or subtracting least significant digits. These people experienced great difficulty with the computerized test and showed large variations in performance. Another problem was that writing or keying in the answers took a significant proportion of the time and confounded numerical facility with motor control. The moral of the story is that computerizing tests should be done with care, and reliability and validity may need to be reassessed.

Most subsequent mathematical processing tests have used addition and/or subtraction of single digits. Chiles et al. (*282*) developed such a test for inclusion in a multiple-task performance battery and used it mainly to assess mental workload and time-sharing capability (*283–285*).

Wanner and Shiner (*286*) used a subtraction test to study working memory. They presented problems, one character at a time, with parentheses on the left, e.g., (5–4)–1, or right, e.g., 5–(4–1) interrupted at intervals by a series of words. Subjects had to solve the problems or recall the words; the authors found that errors were related to the transient memory load imposed by pending operations in the subtraction task. For example, the transient memory load for right-parenthesis problems was greater than that for left-parenthesis problems, since subjects had to defer the subtraction until the whole problem had been presented.

Perez (*287*) used a subtraction and addition test to study working memory and storage. He measured reaction time and accuracy for both algebraic and reverse Polish notation and found that errors were related to confusion between operations, e.g., adding instead of subtracting; reaction time varied with the number of different operations, e.g., + − + was slower than + + +; and after very little practice with the unfamiliar reverse Polish notation, which minimizes transient memory load, performance was better than that using algebraic notation.

Shingledecker (*288*) developed a subtraction/addition task as part of the US Air Force test battery, the criterion task set (CTS); this test was further developed for the STRES battery. The test is presented by computer and consists of several problems, each consisting of three single digits and two operators, e.g., 5 + 4 − 2. The subject has to calculate the answer and say whether it is greater or less than 5 by pressing an appropriate key. This is to minimize the proportion of time taken up by inputting the answer (which is significant in the NF test), and to maintain conformity with the binary responses required by other tests in the STRES battery.

Regarding reliability, the NF test was available in 20 alternate forms, all designed to present similar degrees of difficulty (*289*), although forms 17 and 18 were reported to be significantly harder than the others (*276*). Seales et al. (*290*) studied a test involving addition or subtraction of two three-digit numbers, multiplication of two two-digit numbers, and division of a four-digit number by a two-digit number and reported test–retest correlations of 0.935, 0.941, and 0.921 for total problems attempted, total correct, and correct minus incorrect, respectively. Chiles et al. (*291*) studied a test involving the addition of two two-digit numbers and the subtraction of a third two-digit number and reported test–retest correlations of 0.91 and 0.71 for speed and accuracy, respectively.

Regarding validity, numerical ability has repeatedly been identified as a discrete factor in factor analysis studies of skilled performance (*275*). Construct validity can be illustrated by considering that performance may be broken into four processing stages: *a*) retrieval of arithmetic information from long–term memory; *b*) updating of information in working memory; *c*) sequential execution of arithmetic operations; and *d*) numerical comparisons. Subjects appear to rely not on procedures such as counting, but on a well organized memory structure, storing mathematical tables in their heads (*292–294*). Also, research using multidigit addition (*295*) has shown that complex mathematical problems are solved in series of simple steps in working memory.

Mathematical–processing tests have been widely used to study a variety of stressors such as the effects of wearing protective clothing (*296–299*) and effects of drugs such as alcohol (*300*), atropine and diazepam (*105,301–303*), pyridostigmine (*304*), hyoscine (*305,306*), caffeine and sleep loss (*307*), and exposure to methyl chloride (*308*).

Memory Search

Most memory tests are unsuitable for repeated use owing to task-specific learning effects. Sternberg's memory search test (*175–179*) was chosen because it has been widely used, it is perhaps the most sensitive of the short-term memory tests that are suitable for repeated testing, and like the RT test, it is able to identify the information-processing loci of action of stressor effects.

Sternberg (*175–179*) described a series of studies of memory search processes using the additive factor method. The basic experimental task presents a set of digits (the memory set) followed by a probe digit. Subjects have to say whether or not the probe is a member of the memory set, and their RTs are measured. For example, if the set were 8 3 7 1 and the probe were 3, the correct response would be "yes"; if the probe were 4, the correct response would be "no".

There are three main variations on this basic procedure. The first is the fixed set procedure in which one set is presented and followed by several probes. The second is the varied set procedure in which different sets, each followed by a single probe, are presented on every trial. The third is the mixed set procedure, which is a mixture of the two, e.g., 10 sets each followed by 10 probes. The version of the memory search test in the STRES battery uses a fixed set procedure.

To perform this task correctly, the subject must perform several operations in sequence. First, he must memorize the digit set before presentation of the probe;

otherwise, it will contaminate the reaction time. Recognition and storage of digits (or letters) typically take 250 to 500 msec/item. When the probe is presented, the subject must first detect and recognize it and then perform some sort of search and comparison of the probe with the items held in memory. The outcome of the memory search process provides the information necessary for the subject to select an appropriate response. Thus, the task includes detection, recognition, memory search and comparison, and response selection stages.

Sternberg (*175*) tested the hypothesis that the memory-search stage could be affected by varying the number of items in the memory set. If this is true, then RTs should change with memory-set size, and something might be learned about the memory search process. Two basic memory search strategies that predict different RT functions can be identified: a serial search, which can be self-terminating or exhaustive, and a content-addressable search.

In a serial search, the memory-set items are stored in separate addresses in memory, and the probe is compared successively with the contents of each address. In a serial, self-terminating search, the search stops when a match is found or continues to the end if a match is not found. In this case, when RT is plotted against set size, the slope of the function for "yes" responses will be about half that for "no" responses since, on average, only half the memory set need be searched before a match, if present, is found. (A match must be equally likely in all positions in the set.) In a serial exhaustive search, the search continues to the end whether a match is found or not. In this case, the "yes" and "no" response slopes will be the same.

In a content-addressable search, memory locations are reserved for all members of the population from which the sets are drawn and given the content "no". When the set is presented, the contents of the corresponding items in memory are changed to "yes". For example, if the set is 3 7 2, then the contents of addresses 3, 7, and 2 are changed to "yes". When the probe is presented, the corresponding address is accessed and the answer is immediately available. In this case, changes in set size will not affect memory-search time, and the slope of the RT function will be zero for both "yes" and "no" responses.

Sternberg (*175*) found that RT increased linearly as a function of set size, with the intercept reflecting stimulus encoding, the slope reflecting memory search, and the presence or absence of a match ("yes" or "no" response) reflecting response selection. Further, the "yes" and "no" response slopes were the same. Thus, Sternberg's subjects performed serial exhaustive searches. This conclusion has been generally supported by other investigators, although it has been reported that subjects can change to content-addressable searches with sufficient practice (*309*).

It is probable that, in real life, search strategies vary with the information content of the set items and probes. If asked whether "4" is in an unfamiliar telephone number, most people would perform serial searches. A content-addressable search is unlikely since it is improbable that the contents of the memory location of "4" would contain enough information about whether it is in the telephone number. In contrast, if asked whether butter is in the refrigerator, people do not normally have to search through all the foods kept in memory under refrigerator. Rather, the concept of butter is enough to say whether it is stored in the refrigerator, independently of the number of foods stored.

In another study, Sternberg (*176*) covaried both the memory-set size and the clarity of the probe. On half the trials, the probe digit was presented clearly, and on the other trials it was degraded by placing it behind a masking screen of dots. Logically, it should take longer to recognize a degraded digit than a clear digit. Thus, the RT to degraded digits should be longer than that to clear digits. Further, it is also logical to assume that once the recognition stage has given the probe a label, however easy or difficult it may have been to do so, the rate of memory search will be the same. Thus, the intercept of the reaction time function should change, but the slope should not. Sternberg found that degrading the probe affected only the intercept, indicating that it affected the recognition stage but not the memory-search stage.

This rationale may be applied to other variables. Generally, if task variables have additive main effects, they are inferred to affect separate processing stages. If they have interactive main effects, they are inferred to affect at least one common stage. Thus, an experimental variable such as a drug that interacts with memory-set size is assumed to affect memory search, whereas a variable whose effect is additive to memory-set size is assumed to affect a stage or stages other than memory search.

Many variations on Sternberg's original method have been studied (*179*); the main findings are described below.

Formally, or physically, similar stimuli are scanned more rapidly than stimuli with only associational similarity, and stimuli in the same modality are scanned more rapidly than those in different modalities (*310–312*).

Varying the presentation rate of the memory set items has little or no effect on RT (*313*), but changing the delay between the memory set and the probe affects processing of the memory set; at short delays, memory search and comparison are held up until memory-set processing is complete (*314*). RT is faster when the stimuli are organized, such as in a numerical sequence. RT is also faster on negative trials as a function of numerical separation between the probe and the memory set (*315,316*).

Emphasis on speed or on accuracy each produces strong practice effects on the intercept, but not on the slope, of the RT function (*317*). RT is decreased with increasing delay of a probe after presentation of items that subjects have been told to remove mentally from the memory set (*318,319*). RT decreases as a function of the number of items common to the memory and probe sets (*320–322*).

RTs to pictorial stimuli are faster when processed by the right cerebral hemisphere, and reaction times to letters are faster when processed by the left hemisphere. When stimuli are presented to the slow hemisphere for that type, the intercept of the RT function increases but the comparison rate is unaffected (*323*).

Linear and increasing RT functions have been observed for a wide variety of stimuli, including visual and auditory digits and letters, two- and three-digit numbers, shapes, pictures of faces, drawings of common objects, words of various lengths, colors, and phonemes (*324–330*). The slopes of the RT functions to these types of stimuli differ systematically. The "yes" and "no" functions have been found to remain linear and parallel for memory sets of up to 10 letters (*331*) and up to 12 common words (*332*).

Linear and increasing RT functions have been observed in people of differing personalities, various ages ranging from children to elderly adults, and in normals, alcoholics, schizophrenics, and the brain-damaged mentally retarded. Aging and mental retardation both produce increased slopes compared with those of young, healthy adults (*333,334*). Children of 8 years of age produce RT functions with

higher intercepts but the same slopes as those of young adults (*325,334*). Introverts are slower than extraverts at scanning for semantic features (*335*).

The effects of extended practice vary with the procedure. If the same fixed set is used over many days, then the RT function becomes flatter and negatively accelerated, particularly when the probes are consistently associated with one or other response. There is some evidence that subjects develop a content-addressable search strategy and that processing becomes automatic rather than controlled (*309,336,337*). If the memory sets are changed from trial to trial or from session to session and stimuli are not consistently associated with particular responses, then extended practice affects the intercept but not the slope (*338*).

The Sternberg task has been used widely, in a variety of forms, to identify the information processing loci of action of several drugs.

With industrial chemicals, Smith and Langolf (*339*) found that four levels of exposure to mercury affected the slope but not the intercept of the RT function. Maizlish et al. (*340*) reported that long-term exposure to mixtures of organic solvents had no effect.

With social drugs, Oborne and Rogers (*341*) reported that various combinations of caffeine and alcohol affected the slope but not the intercept of the RT function. Tharp et al. (*342*) reported that alcohol impaired response selection. Roth et al. (*37*) found that alcohol and marijuana differed significantly from placebo but not from each other and that marijuana increased overall RT.

With benzodiazepines, Subhan (*343*) reported that flunitrazepam and triazolam impaired stimulus encoding and serial comparison stages, whereas lormetazepam had little or no effect. Rizzuto (*344*) found that 5 mg diazepam did not affect performance, whereas 10 mg increased RT but did not affect error scores.

With hypnotics, Rundell et al. (*345*) and Williams et al. (*346*) reported that secobarbital affected stimulus encoding, but Mohs et al. (*347*) reported that it had no effect. With antidepressants, McNair et al. (*348*) found that amitryptiline improved RT generally, whereas amoxapine had no effect.

With stimulants, Naylor et al. (*349*) reported that methylphenidate speeded response selection but not stimulus evaluation, and Mohs et al. (*347*) reported that methamphetamine had no effect. With anticholinesterases, Wetherell (*350*) found that physostigmine (previously reported to improve memory) improved stimulus recognition but not memory search. With hormones, Ward et al. (*351*) reported that melanocyte-stimulating hormone and adrenocorticotrophic hormone improved stimulus encoding but did not affect memory search in men or women.

With regard to other stressors, Lorenz and Lorenz (unpublished data) reported that both speed and accuracy were impaired during simulated diving to 560 m using heliox and to 360 m using trimix (5% nitrogen). Briggs et al. (*352*) reported that concurrent tracking affected the intercept of the RT function but not the slope, and Crosby and Parkinson (*353*) reported that performance of a ground-controlled approach by pilots had a similar effect. Wetherell (*354*) reported that car driving affected the intercept but not the slope for "yes" responses and both the intercept and the slope for "no" responses. He suggested that subjects were less certain about a "no" than about a "yes" decision and carried out more searches to accumulate confidence before responding.

Some investigators have studied the relationship between the memory search task and the P300 component of the evoked cortical potential, which has been claimed to reflect information processing activity. Gomer et al. (*355*) reported that the P300 was enhanced for "yes" responses and that the difference in P300 between "yes" and "no" responses increased with memory-set size. Brookhuis et al. (*356*) reported that the memory search task results indicated a self-terminating search whereas the P300 results indicated an exhaustive search. Adam and Collins (*357*) reported that P300 latencies increased with memory-set size up to seven digits. Ford et al. (*358*) reported that RT was slower in older subjects than in younger subjects, but there was no difference in P300 latency or amplitude. However, Pfefferbaum et al. (*359*) reported that the P300 amplitude increased with memory-set size and decreased with age.

The reliability of the memory search test RTs has been reported as generally greater than 0.70 (*360,361*). Split-half reliabilities were measured by Boer (*362*) for slopes and intercepts for memory set sizes of one and four letters (Table 3).

The validity of the test is related to that of the additive factor method, as described above for the RT test.

Spatial Processing

Factors relating to spatial ability have been reported in factor analyses for some time (*363–369*). More recently, Lohman (*370*) suggested that there existed a broadly defined spatial factor with several subfactors: spatial relations, spatial orientation, and visualization. Spatial ability has also been studied using the information-processing approach (*371–373*).

The spatial processing test used in the STRES battery consists of pairs of four-bar histograms with the two histograms in each pair presented sequentially. The first histogram of each pair is presented upright and the second rotated through either 90° or 270°. For each pair, subjects have to say whether the second histogram is the same as or different from the first by pressing appropriate response keys.

The test is based on that used in the CTS (*288*), which in turn was based on an earlier task devised by Fitts et al. (*374*) and used by Chiles et al. (*282*) to study work schedules and long-term isolation. The test taps the visualization ability involved in mental reorientation, perceptual speed, and probably closure speed (*370*).

Kennedy et al. (*375*) reported that the Fitts et al. (*374*) histogram test has a test–retest reliability of 0.90. Chiles et al. (*282*) reported that their version of the test has a split-half reliability of 0.75; Schlegel and Gilliland (*307*) reported that the STRES version of the test has a reliability coefficient of 0.67.

Regarding validity, Kennedy et al. (*375*) reported that the Fitts et al. (*374*) histogram test gave a correlation of 0.71 with Klein and Armitage's (*376*) pattern comparison test, and the histogram scores loaded on to the same factor as other tests with spatial components such as the manikin test (related to Lohman's spatial orientation factor), code substitution, and the Klein and Armitage pattern comparison test (both related to Lohman's spatial relations factor). The histogram test also loads on to a motor control factor, perhaps because it was originally used in a paper-and-pencil version.

Regarding sensitivity to stressors, there is little information as yet. Rizzuto (*344*) reported that 10 mg diazepam increased reaction times but did not affect accuracy. Some sensitivity may be inferred from

Table 3. Split-half reliability coefficients for the memory search test.

Memory set size	Slope	Intercept
1	0.32	0.74
4	0.62	0.65
1 and 4 combined	0.76	0.87

findings using tests with which the histogram test is correlated. For example, the manikin test is sensitive to the effects of deep diving (*377,378*), the pattern comparison test is sensitive to cyclical variations in arousal (*376*), and the Fitts et al. (*374*) histogram test is sensitive to the effects of long-term isolation (*282,284*).

Unstable Tracking

Many tracking tests exist; they are classified into pursuit (where the subject pursues a moving target), compensatory (where the target remains stationary and the tracking cursor drifts), and combined compensatory/pursuit. Some adaptive tests also exist in which the evasive movements of the target increase as the tracking cursor approaches. A further classification is in terms of the order: zero-order is when the response device (e.g., joystick) controls just the cursor direction; first-order is when the joystick movement controls direction and speed; second-order (rare) is when the joystick controls acceleration.

The unstable tracking test used in the STRES battery involves using a joystick to keep a cursor aligned next to a target positioned in the middle of the monitor screen. The cursor starts aligned, but it will accelerate away, left or right; the further it moves, the greater the joystick movement has to be to correct it. If the cursor reaches a boundary, it will reappear at the target and begin moving away again. The test is analogous to balancing a billiard cue (*379*): if it moves only slightly, then only fine corrections are needed, but the further it falls, the more it accelerates and the greater the correction has to be.

The test was inspired by analyses of aircraft handling (*380,381*) and developed by Jex et al. (*382*), based on Fourier analysis and linear feedback control theory. Tracking performance can be described by the linear differential equations, or transfer functions, incorporated into a quasilinear class model of the human operator. In such models, the response to tracking input signals, although nonlinear, is approximated by a linear transfer function called the describing function and a separate, nonlinear component called the remnant. The parameters of quasilinear models, e.g., time delay and gain, appear to correspond to specific characteristics of human control behavior in man–machine systems. For example, the time delay appears to be analogous to discrete reaction time (*138*). The unstable tracking test is based on the quasilinear crossover model of McRuer and Jex (*383*) and used in the CTS (*288,382*).

Damos et al. (*384*) reported that unstable tracking performance stabilized after 10 sessions and had a test–retest correlation of 0.764. However, Damos et al. (*385*) later reported that performance became stable after 105 brief practice trials, although performance slowly improved over 14 days. Schlegel and Gilliland (*307*) reported that the STRES version had a reliability coefficient of 0.83 for mean absolute error and 0.82 for control losses. The validity of the test is based on the findings that human performance closely follows theoretical assumptions (*382*).

The test has proved to be sensitive to a variety of stressors including alcohol (*386,387*), carbon monoxide (*388*), diazepam (*344*), sleep loss (*307*), saturation diving (Lorenz and Lorenz, unpublished data), and variations in acceleration force (*389–391*).

Grammatical Reasoning

Several logical reasoning tests have been proposed. For example, Wason (*392*) used sentences describing a number as odd or even, such as "seventy-six is an even number", or "sixty-two is not an even number", and found that negative statements were verified more slowly than were affirmative statements. He suggested that this was due to the extra time required to invert the negative form (e.g., "not even") to the affirmative (e.g., "odd").

Slobin (*393*) used sentences followed by pictures (e.g., a cat chasing a dog), and asked subjects to say whether the sentence correctly described the picture. Chase and Clark (*394*) and Clark and Chase (*395,396*) used sentences describing the relationship of * and + symbols, e.g., "the star is not above the plus." These authors also found that negative sentences took longer to verify than did affirmative ones, presumably because subjects inverted them.

Baddeley (*1*) described a test of verbal reasoning based on grammatical transformation, which has since become perhaps the best known and widely used of its type. The test consists of sentences, each followed by two letters: AB or BA. The sentences described the order of the two letters, and subjects had to say whether the description was true or false. Examples include A follows B—AB; B does not follow A—BA; A is preceded by B—BA. Thirty-two different problems can be generated by combining the verb follow or precede, the active or passive voice, affirmative or negative construction, the order of A and B in the statement, and the order of A and B in the letter pair. Baddeley (*1*) found that affirmative sentences were verified faster than negative ones and active sentences verified faster than passive.

The version of the test used in the STRES battery was derived from Shingledecker (*288*), who substituted the symbols used by Clark and Chase (*395,396*) for the letters used by Baddeley (*1*) to avoid any bias from the natural, alphabetic order of A and B. The STRES version was further modified to avoid cultural and language bias. In particular, passive sentences were omitted since the passive voice is seldom used in some languages, e.g., German. Since this left only the active, and hence easier, problems, it was decided to use two statements describing the order of three symbols. The symbols were &, #, and *, and the statements consisted of a symbol, the word BEFORE or AFTER, and a second symbol, e.g.,

& AFTER #	# BEFORE *	* AFTER &
* BEFORE &	# AFTER &	# BEFORE *
*&#	&#*	#*&

If both statements have the same truth value, i.e., both true or both false, the subject responds "same;" if the statements have different truth values, i.e., one true and one false, the subject responds "different."

Regarding reliability, Baddeley (*1*) reported a test–retest reliability of 0.80 for the original paper and pencil form of the test. Carter et al. (*397*) studied a similar test, but of a 1-min duration, and reported a test–retest correlation of 0.82. Schlegel and Gilliland (*307*) studied the CTS version and reported a reliability coefficient of 0.83.

Regarding validity, Baddeley (*1*) reported a correlation of 0.59 with the British Army Verbal Intelligence Test, Carter et al. (*397*) reported a correlation of 0.44 with the Wonderlic Test of Mental Ability, and Wetherell (*398*) reported a nonsignificant correlation of 0.22 with Raven's Standard Progressive Matrices. Further support that the test taps verbal ability comes from the findings that recall of verbal memory loads is impaired by the grammatical reasoning test but not by spatial reasoning tests: the test taps the verbal articulatory loop and not the visuospatial scratch-pad of working memory (*399,400*).

Perhaps the best known use of the test was in the early studies of working memory: concurrent memory loads of six letters slowed grammatical reasoning performance but did not affect accuracy (*401,402*).

The test is reported to be sensitive to the effects of diazepam (*105*), atropine (*304*), physostigmine (*403*), car driving (*250*), nitrogen narcosis (*404*), fatigue (*405*), and anxiety before decompression (*406*) but not to anxiety before simulated deep sea diving [Lorenz and Lorenz, unpublished data; (*377*)].

Recently, Salame (*407*) commented that the STRES version of the test could be performed without reference to the symbols as long as three rules were followed: *a*) when there is only one match, the correct answer is always "same;" *b*) when there are two matches, the correct answer is always "different;" and *c*) when there is no match, the correct answer is always "different." Salame suggested that subjects could discover these rules and that the test would no longer measure logical reasoning. While most subjects probably would not have the opportunity or ability to analyze the structure of the test in this way, even if they did, the test could still be considered to measure a form of logical reasoning. Nevertheless, Salame's criticism and proposed improvements should be borne in mind.

Dual Task

The dual task is a combination of the unstable tracking task and the memory search task. The memory set is presented and tracking begins when the subject presses a key to say that he has memorized the set. Probe items are presented immediately above the central tracking target, and the two tests run concurrently although separate scores are recorded for each as described above.

The reliability of each of the tests is discussed above. There is no direct evidence concerning their reliability in combination. However, the test–retest reliability of tracking with other concurrent tasks is reported to be fairly good (*408*).

Regarding validity, there have been some attempts to identify a general time-sharing factor but with inconclusive results (*408–410*). The differences between single-task performance and dual-task performance on these tasks become smaller with practice (*151*), but it is not certain whether this reflects improved time-sharing ability or reduced resource demands.

The sensitivity of the two tests separately was described above. The few investigations of tracking with concurrent memory search so far carried out have been concerned primarily with the development of multiple resource models of performance. However, tracking combined with other tasks has proved sensitive to several stressors including alcohol, caffeine, and methyl chloride (*411*), carbon monoxide and methylene chloride (*412*), G-stress (*413*), and sleep loss (*414*).

Study Design

It is not intended here to present a treatise on study design; detailed information can be found in most textbooks on experimental and applied psychology. However, some practical comments might help set the rest of the paper into some context.

Most psychological tests are used to determine an individual's score with respect to normative data for purposes of diagnosis or classification. Performance tests are used differently: there are usually no norms, and comparison, reference, or control data must be obtained from other sources. Usually it is obtained concurrently with test data, but sometimes it can be obtained from previously gathered data. In any case, performance tests are, or should be, administered as part of a carefully designed procedure.

An experimental design systematically manipulates independent variables to discover their effects on dependent variables. To attribute cause and effect correctly, all other variables must be controlled, usually by eliminating those that can be eliminated, counterbalancing those that cannot, or measuring those that cannot be eliminated or counterbalanced. Variables that are not accounted for can confound the results, i.e., make it impossible to tell which variable caused which effect.

Experimental studies of the effects of drugs and other environmental stressors on performance tend to follow two general types of design. The first is between subjects, or parallel groups, in which subjects are allocated randomly to groups, and each group receives one of the treatments. The main advantage is flexibility in that the loss of one subject's results can easily be repaired by adding more subjects. The main disadvantage is that a statistically viable number of subjects is needed for each group, and can run to large numbers.

Randomization is assumed to distribute subject characteristics evenly between the groups but bias can occur, and there is a risk of confounding treatment effects with those arising from differences between the groups. For example, by chance, one group may be more resistant to drug effects or may be better at performing the tests. It is possible to match the subjects in terms of known characteristics, but doing so might unmatch them in terms of unknown, but important, characteristics. Some investigators argue that subjects should be assigned to treatment groups completely at random; others argue that doing so might introduce bias, e.g., studying groups in sequence could cause confounding with seasonal changes in a long study. Thus, it might be better to build up all the groups at the same rate.

The second type of design is within-subjects or repeated measures, in which all subjects receive all treatments, including control treatments. The treatment administration orders should be randomized or counterbalanced, and in cases where only two treatments are involved, e.g., a drug and a placebo, it is usual for half the subjects to receive the placebo first and the other half to receive the active treatment first. This is sometimes called a crossover design. In cases where more than two treatments are involved, more complex counterbalancing is needed, e.g., factorial designs, or Latin Squares. Here, the investigator must give the treatments in such a way that they are not confounded with the temporal order. Some examples covering four treatments are shown in Table 4. Example 1 is not very good because some sequences are repeated: if there is a hangover effect of, e.g., drug A on drug B, then subjects 1, 3, and 4 will show a bias. Example 2 is better; here, all sequences are completely counterbalanced.

With some numbers of treatments, e.g., three, the design cannot be so economical because the orders cannot be completely counterbalanced. Sometimes the only recourse is to study all combinations of orders—six for three treatments. Only very brave investigators attempt more than four treatments.

The advantages of within-subject designs derive mainly from the fact that each subject acts as his own control; thus, fewer subjects are needed than for a between-subjects design. Since each subject acts as his own control, there is no confounding of treatment effects with differences between subjects. One disadvantage is inflexibility: the loss of one subject's

Table 4. Examples of Latin Squares.

	Example 1				Example 2			
	Order				Order			
Subjects	1st	2nd	3rd	4th	1st	2nd	3rd	4th
1	A	B	C	D	A	B	C	D
2	B	C	D	A	B	D	A	C
3	C	D	A	B	C	A	D	B
4	D	A	B	C	D	C	B	A

results for only one treatment means that all of that subject's treatments have to be replaced, and the subjects must be studied in certain multiples, e.g., a 4×4 Latin Square requires subjects in multiples of four. Another disadvantage is that treatment effects can be confounded with the effects of repeated testing such as fatigue and learning, and other intercurrent events may occur in one treatment but not another, e.g., paydays, changes in health, emotions, and biological rhythms. Such factors lead to asymmetric transfer and other effects (*182,415*) that can confound treatment effects.

Sometimes it is not possible to use a within-subjects design or even to study a concurrent control group, e.g., in measuring the performance of people who have been accidentally exposed to a stressor such as a toxic chemical or who have been exposed over a long period of time. Here, control groups must be sought, e.g., from among those in the same job who were not exposed, or if this is not possible, from people in similar jobs. This raises many problems since the control group would have been subjected to different influences, and it might be impossible to tell whether any effects are actually due to the accidental exposure.

In any case, the choice of experimental design is not normally the investigator's alone; it is often forced by circumstances or, at best, is a compromise between what the investigator considers best, the sponsor's demands, and the time, subjects, and finances available.

A word on placebos is appropriate. A placebo should match the active treatment(s) in all respects except that it does not contain any active agent. This is actually a tall order; it can often be quite difficult to match size, shape, color, weight, taste (or other sensory aspect), sometimes solubility, as well as means and style of administration. An example is alcohol, the taste and smell of which cannot easily be disguised or mimicked, and which has proved a considerable problem in experimental studies. Some authors have had to produce quite complex cocktails of such ingredients as orange juice, peppermint, herbs, and spices. The taste and smell can be reduced by dilution, but this might mean that subjects have to drink prohibitive amounts of liquid to get the required dose.

Just as placebos are necessary for experimental control, some investigators also use an active, or positive, control, sometimes called a verum. This is a treatment intended to produce effects and is often used in cases where the investigator suspects that the treatments being tested might have little or no effect. Given that an experiment is designed to test the null hypothesis (that there is no effect), it is impossible to prove that there is no effect; investigators can only say that they have not shown an effect. Experimental and applied psychologists learn from an early age that absence of evidence is not evidence of absence. There may be several reasons for this, the main one being that the tests used might have been insufficiently sensitive. This is often difficult for sponsors to understand or accept since it is often the finding that they want, and no investigator likes to admit that his tests were not good enough. A verum, which shows an effect, and hence, that the tests were good enough to detect it, can lend support to a conclusion that the test treatment actually has no effect. Vera are also useful to distract subjects.

All treatments—test, placebo, vera—should be administered double-blind, i.e., neither the subjects nor the experimenters know which treatment is which. The identity is known only to a third party who takes no role in the study other than preparing, allocating, and recording treatments. The third party certainly does not administer tests or even meet the subjects, and it is helpful if he is not on speaking terms with the experimenters. This is also sometimes difficult to achieve since some drugs have characteristic effects that quickly become apparent to the experimenters and sometimes to the subjects. For example, atropine causes the pupils of the eye to enlarge (mydriasis), which can easily be seen by experimenters and also causes blurring of vision by cycloplegia, which can be detected by the subjects.

One point that is often forgotten is that double-blind procedures should extend all the way through a study, including the analysis of results. It is too easy to say that, for example, computer-presented and scored tests are not subject to bias, but some scores may have to be interpreted during preliminary data reduction or missing data may have to be estimated. The purists would say that this should not happen, but in the real world it does. It is quite possible to analyze results with treatment groups identified by codes: final identification should be left until all the significant differences have been calculated.

A word on learning is also appropriate. Learning is sought after, e.g., by educational psychologists, but experimental and applied psychologists would rather it did not exist. All investigators give, or should give, subjects at least some familiarization on the tests beforehand. Many investigators train their subjects and often report that they did so until the subjects reached a plateau of performance. This sounds reasonable but it is misleading. Wetherell (*280*) noted that the number of Moran and Mefferd (*275*) number facility tests needed to achieve a performance plateau increased from about three to over 60 with successive reports of the test's properties and carried out his own studies. He found that subjects always achieved a plateau of 15 to 20% improvement no matter how much training they were given, as long as they knew beforehand how much time they would have. When further training was given unexpectedly, subjects improved by another 10 to 15%.

Rabbitt (*416*) showed that performance improvements can occur over prolonged periods, i.e., months, even with a measure such as simple reaction time, which is often thought to involve minimal learning. Bittner et al. (*45*), in their studies of tests for the PETER battery, reported that while performance achieved some stability over several training sessions, it was still increasing. Thus, Parrott (*192*) cautions "beware of any study blithely reporting the absence of learning effects," no matter what the authors claim.

Learning is a particular problem with memory tests since by definition they require subjects to learn things. The problem lies mainly in that subjects do not forget them; items memorized from a word list in one part of a study can intrude and affect memory for other items that need to be memorized in a later part. The problem is less serious if the items are alphanumeric characters because they can usually be randomized, as in tests such as Wechsler's digit span (*50*), but words have more meaning and are less easily forgotten. Such effects are sometimes studied in themselves, but most investigators studying stress effects would rather do without them.

Some other points should also be considered with respect to studies of the effects of drugs. The first is that most drugs are taken to help cure or alleviate illness, and they will have different effects on patients than on healthy experimental subjects. Employing patients as experimental subjects does not help; the effects of the drug would be confounded with those of the illness, and there are problems in defining

the type and level of illness that constitutes a patient, especially with drugs having subtle effects such as tranquillization.

A distinction must also be drawn between the acute effects of a single dose of a drug given to drug-naive experimental subjects and the chronic effects of repeated dosing more often used in real life. However, it is well known that many patients tend to modify their prescribed dosing regimens to suit their way of life and may periodically suffer from effects similar to those experienced by first-time users.

In choosing tests, several factors must be considered. The first is the range of skills or psychological functions that needs to be covered. Sometimes there is only one and the investigator may be able to measure performance on the actual task or some simulation or suitable laboratory test. More often, investigators are required to make statements on a wide range of skills; here, tests should be chosen according to some model or taxonomy.

Second, tests are often chosen to cover the likely effects of the stressor, based on prior evidence or some rationale. For example, in psychopharmacology, tests are chosen to cover the range of behavior that could result from the pharmacological actions of a drug. Performance effects are most likely to arise from known drug effects on the central nervous system, but they may also arise indirectly through subjects' awareness of peripheral effects, e.g., on the eye or on heart rate or force, which can be especially alarming. Thus, it is premature to assume that a drug is behaviorally inactive simply from pharmacological evidence, which may itself be questionable, that it does not cross the so-called blood–brain barrier. This barrier is not as final as it sounds; it allows, and often facilitates, the passage of drugs and metabolites into the brain, and its selective permeability and topographical distribution vary between and within species and within the same animal according to age, state of health, etc. (*417,418*).

Also, it cannot be assumed that behavioral effects of a drug are governed by its pharmacokinetics. Rates of drug absorption and initial distribution might determine the onset of behavioral effects, but their severity and duration depend less on metabolism and excretion than on the subject's ability to adapt and compensate or surrender. These factors depend in turn on age, arousal level, reserve capacity, personality, mood, sex, self-image, and time of day, week, month, or even year.

Third, all tests should be sensitive, reliable, and valid, but it is not always possible to achieve all three criteria, especially when increasing reliability or validity leads to a decrease in sensitivity. In such cases, it is often useful to include a test of proven sensitivity, even though its reliability and validity are unknown, in order to show that some effects can be detected by some means and to lend support to a conclusion that there is no effect.

A major problem in any study is to obtain enough subjects with the requisite qualifications, (e.g., age, sex, degree of stressor and test naivety) to form a statistically reliable sample representative of the population for whose benefit the work is intended. Subjects are always scarce and, since for most studies they must be volunteers, some bias must be tolerated. Some groups of people are notoriously reluctant to volunteer for anything, while others can seem rather too eager, perhaps from subtle pressures or dubious motives that could have more effect than the drug on the behavior under test.

REFERENCES

1. Baddeley AD. A 3-minute reasoning test based on grammatical transformation. Psychon Sci 10:34 (1968).
2. Cull C, Trimble MR. Automated testing and psychopharmacology. In: Human Psychopharmacology: Materials and Methods, Vol 1 (Hindmarch I, Stonier PD, eds). Chichester, UK:John Wiley & Sons, 1987.
3. Wittenborn JR. Psychomotor tests in psychopharmacology. In: Human Psychopharmacology: Materials and Methods, Vol 1 (Hindmarch I, Stonier PD, eds). Chichester, UK:John Wiley & Sons, 1987.
4. Hindmarch I. Psychomotor function and psychoactive drugs. Br J Clin Pharmacology 10:189–209 (1980).
5. Adams RG. Pre-sleep ingestion of two hypnotic drugs and subsequent performance. Psychopharmacology 40:185–190 (1974).
6. Malpas A, Rowan AJ, Joyce CRB, Scott DF. Persistent behavioural and electroencephalographic changes after single doses of nitrazepam and amylobarbitone sodium. Br Med J 2:762–764 (1970).
7. Ashton H, Hall, EH, Savage RD, Telford R, Thompson JW. A small controlled study to determine the time of onset of action of oxypertine after oral administration in normal subjects. Postgrad Med J (September):14–18 (1972).
8. Borland RG, Nicholson AN. Immediate effects on human performance of a 1,5-benzodiazepine (clobazam) compared with the 1,4-benzodiazepines chlordiazepoxide hydrochloride and diazepam. Br J Clin Pharmacol 2:215–222 (1974).
9. Gagné RM, Fleishman EA. Psychology and Human Performance. New York:Holt, 1959.
10. Fleishman EA, Hempel WE. Factorial analysis of complex psychomotor performance and related skills. J Appl Psychol 40:96–104 (1956).
11. File SE, Bond AJ. Impaired performance and sedation after a single dose of lorazepam. Psychopharmacology 66:309–313 (1979).
12. Van Houten PL, Zenhausern R. Meprobamate and absolute auditory thresholds. J Aud Res 7:253–257 (1967).
13. Jones O. Relationship between visual and auditory discrimination and anxiety level. J Gen Psychol 59:111–118 (1958).
14. Bernstein ME, Hughes FW, Forney RB. The influence of a new chlordiazepoxide analogue on human mental and motor performance. J Clin Pharmacol 7:330–335 (1967).
15. Bond AJ, Lader MH. The residual effects of flurazepam. Psychopharmacology 32:223–235 (1973).
16. Hindmarch I. A 1,4-benzodiazepine, temazepam (K3917): its effect on some psychological parameters of sleep and behaviour. Arzneim Forsch 25:1836–1839 (1975).
17. Veldkamp W, Straw RN, Metzler CM, Demissianos HV. Efficiency and residual effect evaluation of a new hypnotic, triazolam. J Clin Pharmacol 14:102–111 (1974).
18. Malpas A, Joyce CRB. Effects of nitrazepam, amylobarbitone and placebo on some perceptual, motor and cognitive tasks in normal subjects. Psychopharmacology 14:167–177 (1969).
19. Davies KL, Hollister LE, Overall J, Johnson A, Train K. Physostigmine: effects on cognition and affect in normal subjects. Psychopharmacology 51:23–27 (1976).
20. Ghoneim MM, Mewaldt SP. Studies on human memory: the interactions of diazepam, scopolamine and physostigmine. Psychopharmacology 52:1–6 (1977).
21. Peck AW, Adams R, Bye C, Wilkinson RT. Residual effects of hypnotic drugs: evidence for individual differences on vigilance. Psychopharmacology 47:213–216 (1976).
22. Adamson GT, Finlay SE. A comparison of the effects of varying dose levels of oxypertine on mood and physical performance of trained athletes. Br J Psychiat 112:1177–1180 (1966).

23. Holmberg GG, William-Ollson U. The effect of benzquinamide, in comparison with chlordiazepoxide and placebo, on performance in some psychological tests. Psychopharmacology 4:402–417 (1963).
24. Wittenborn JR, Flaherty CF, McGough WLE, Nash RJ. Psychomotor changes during the initial day of benzodiazepine medication. Br J Clin Pharmacol 8:69S–76S (1979).
25. Bond AJ, Lader MH. Residual effects of hypnotics. Psychopharmacology 23:117–132 (1972).
26. Lahtinen U, Lahtinen A, Pekkola P. The effect of nitrazepam on manual skill, grip strength and reaction time with special reference to subjective evaluation of effects on sleep. Acta Pharmacol Toxicol 42:130–134 (1978).
27. Idestrom CM, Cadenius B. Chlordiazepoxide, dipiperon and amobarbital: dose effect studies. Psychopharmacology 4:235–246 (1963).
28. Aschoff JC, Becker W, Weinert D. Computer analysis of eye movements: evaluation of the state of alertness and vigilance after sulpiride medication. J Pharmacol Clin 11:93–97 (1975).
29. Zimmermann-Tansella C, Tansella M, Lader M. The effects of chlordesmethyl diazepam on behavioural performance and subjective judgment in normal subjects. J Clin Pharmacol 16:481–488 (1976).
30. Masuda M, Bakker CB. Personality, catecholamine metabolism and psychophysiological response to diazepam. J Psychiat Res 4:221–234 (1966).
31. Church MW, Johnson LC. Mood and performance of poor sleepers during repeated use of flurazepam. Psychopharmacology 61:309–316 (1979).
32. Stitt FW, Latour R, Frane JW. A clinical study of naproxen-diazepam drug interaction on tests of mood and attention. Curr Ther Res 21: 149–156 (1977).
33. Hedges A, Turner P, Harry TV. Preliminary studies on the central nervous effects of lorazepam, a new benzodiazepine. J Clin Pharmacol 16:423–427 (1971).
34. Gendreau P, Sherlock D, Parsons T, McLean R, Scott GD, Suboske MD. Effects of methamphetamine on well-practiced discrimination conditioning of the eyelid response. Psychopharmacology 25:112–116 (1972).
35. Wittenborn JR, Flaherty CF, Hamilton LW, Schiffman HR, McGough WE. The effect of minor tranquilizers on psychomotor performance. Psychopharmacology 47:281–286 (1976).
36. Hindmarch I, Parrott AC. The effect of repeated nocturnal doses of clobazam, dipotassium chlorazepate and placebo on subjective ratings of sleep and early morning behaviour, and objective measures of arousal, psychomotor performance and anxiety. Br J Clin Pharmacol 8:235–239 (1979).
37. Roth T, Kramer M, Lutz T. The effects of hypnotics on sleep, performance and subjective state. Drugs Exp Clin Res 1:279–286 (1977).
38. Hindmarch I, Parrott AC. The effect of a sub-chronic administration of three dose levels of a 1,5-benzodiazepine derivative, clobazam, on subjective aspects of sleep and assessments of psychomotor performance the morning following night time administration. Arzneim Forsch 28:2169–2172 (1978).
39. Bond AJ, Lader MH. Residual effects of flunitrazepam. Br J Clin Pharmacol 2:143–150 (1975).
40. Jones DM, Lewis MJ, Spriggs TLB. The effects of low doses of diazepam on human performance in group administered tasks. Br J Clin Pharmacol 6:333–337 (1978).
41. Jones DM, Jones MEL, Lewis MJ, Spriggs TLB. Drugs and human memory: effects of low doses of nitrazepam and hyoscine on retention. Br J Clin Pharmacol 7:479–483 (1979).
42. Hindmarch I, Clyde CA. The effects of triazolam and nitrazepam on sleep quality, morning vigilance and psychomotor performance. Arzneim Forsch 30:1163–1166 (1980).
43. Busch von M, Klapproth HE, Lücker, Schmitz H. Ein neues psychometrisches Testmodell auf der Basis des "tailored testing" zur ökonomischen und effizienten Erfassung psychopharmakologischer Effekt. Arzneim Forsch 29:859–863 (1979).
44. Kennedy RS, Bittner JR, Harbeson M, Jones MB. Television computer games: a new look in performance testing. Aviat Space Environ Med 53:49–53 (1982).
45. Bittner AC, Carter RC, Kennedy RS, Harbeson MM, Krause M. Performance evaluation tests for environmental research (PETER):evaluation of 114 measures. Percept Mot Skills 63:683–708 (1986).
46. Wiker SF, Kennedy RS. Pepper RL. Development of performance evaluation test for environmental research (PETER): navigational plotting. Aviat Space Environ Med 54:144–149 (1983).
47. Tomlinson L, Andrews D, Merrifield E, Reynolds EH. The effects of antiepileptic drugs on cognitive and motor functions. Br J Clin Practice 18:177–183 (1982).
48. Morris RG, Evenden JL, Sahakian BJ, Robbins T. Computer-aided assessment of dementia: comparative studies of neuropsychological deficits in Alzheimer-type dementia and Parkinson's disease. In: Cognitive Neurochemistry (Stahl SM, Iversen SD, Goodman EC, eds). Oxford:Oxford University Press, 1987;21–36.
49. Warburton D. Drugs and the processing of information. In: Cognitive Neurochemistry (Stahl SM, Iversen SD, Goodman EC, eds). Oxford:Oxford University Press, 1987;111–134.
50. Wechsler D. A standardised memory scale for clinical use. J Psychol 19:87–95 (1945).
51. Wechsler D. A Manual for the Wechsler Adult Intelligence Scale. New York:Psychological Corporation, 1955.
52. Hart J, Hill HM, Bye CE, Wilkinson RT, Peck AW. The effects of low doses of amylobarbitone sodium and diazepam on human performance. Br J Clin Pharmacol 3:289–298 (1976).
53. Shaffer JW, Freinek WR, Wolf S, Foxwell NH, Kurland AA. A controlled evaluation of chlordiazepoxide in the treatment of convalescing alcoholics. J Nerv Ment Dis 137:494–507 (1963).
54. Besser GM, Steinberg H. L'interaction du chlordiazepoxide et du dexamphetamine chez l'homme. Therapie 22:977–990 (1967).
55. Jaattela A, Mannisto P, Paatero H, Tuomisto J. The effects of diazepam or diphenhydramine on healthy human subjects. Psychopharmacology 21:202–211 (1971).
56. Shira RB. A technique for investigating the intensity and duration of human psychomotor impairment after intravenous diazepam. Oral Surg 45:493–502 (1978).
57. Wittenborn JR. Contrasts in anti-depressant medication. Br J Clin Pharmacol 4:153S–156S (1977).
58. Walters AJ, Lader MH. Hangover effects of hypnotics in man. Nature 229:637–638 (1971).
59. Malpas A. Subjective and objective effects of nitrazepam and amylobarbitone sodium in normal human beings. Psychopharmacology 27:373–378 (1972).
60. Peck AW, Bye CE, Claridge R. Differences between light and sound sleepers in the residual effects of nitrazepam. Br J Clin Pharmacol 4:101–108 (1977).
61. Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol 18:643–662 (1935).
62. Uherik A. Methodological problems of psychophysiological research on human subjects. Studia Psychol 15:213–228 (1973).
63. Shor RE. Symbol processing speed differences and symbol interference effects in a variety of concept domains. J Gen Psychol 85:187–205 (1971).
64. Schiller PH. Developmental study of color-word interference. J Exp Psychol 72:105–108 (1966).
65. Dyer FN. The Stroop phenomenon and its use in the study of perceptual, cognitive and response processes. Mem Cognit 1:106–120 (1973).
66. Roden S, Harvey P, Mitchard M. The influence of alcohol on the persistent effects on human performance of the hypnotics Mandrax and nitrazepam. Int J Clin Pharmacol 15:350–355 (1977).
67. Nakano S, Gillespie HK, Hollister LE. A model for the evaluation of anti-anxiety drugs with the use of experimentally induced stress: comparison of nabilone and diazepam. Clin

Pharmacol Ther 23:54–62 (1978).
68. Janke W, Debus G. Experimental studies on antianxiety agents with normal subjects: methodological considerations and review of main effects. In: Psychopharmacology: A Review of Progress 1957–1967 (Efron D, ed). Public Health Service Publication 1836. Washington:Department of Health Education and Welfare, 1968.
69. Latz A. Cognitive test performance of normal human adults under the influence of psychopharmacological agents: a review. In: Psychopharmacology: A Review of Progress 1957–1967 (Efron D, ed). Public Health Service Publication 1836. Washington:Department of Health, Education and Welfare, 1968.
70. McNair DM. Anti-anxiety drugs and human performance. Arch Gen Psychiat 29:611–617 (1973).
71. Kleinknecht RA, Donaldson D. A review of the effects of diazepam on cognitive and psychomotor performance. J Nerv Ment Dis 161:399–411 (1975).
72. Clayton AB. The effects of psychotropic drugs upon driving-related skills. Hum Factors 18:241–252 (1976).
73. Wittenborn JR. Effects of benzodiazepines on psychomotor performance. Br J Clin Pharmacol 7:61S–67S (1979).
74. Seppala T, Korttila K, Hakkinen S, Linnoila M. Residual effects and skills related to driving after a single oral administration of diazepam, medazepam, or lorazepam. Br J Clin Pharmacol 3:831–841 (1976).
75. Dixon RA, Thornton JA. Tests of recovery from anaesthesia and sedation: intravenous diazepam in dentistry. Br J Anaesth 45:207–215 (1973).
76. Haffner JFW, Mørland J, Setekleiv J, Stromsaether CE, Danielsen A, Frivik PT, Dybing F. Mental and psychomotor effects of diazepam and alcohol. Acta Pharmacol Toxicol 32:161–178 (1973).
77. Mørland J, Setekleiv J, Haffner JFW, Stromsaether CE, Danielsen A, Holstwethe G. Combined effects of diazepam and ethanol on mental and psychomotor performance. Acta Pharmacol Toxicol 34:5–15 (1974).
78. Effects of drugs on driving: driving simulator tests of diazepam and secobarbital. Proc Human Factors Soc 1978;259–262.
79. Lawton MP, Cahn B. The effects of diazepam (valium) and alcohol on psychomotor performance. J Nerv Ment Dis 136:550–554 (1963).
80. Clark PRF, Eccersly PS, Frisby JP, Thornton JA. The amnesic effect of diazepam (valium). Br J Anaesth 42:690–697 (1970).
81. Dixon RA, Hatt SD. Controlled clinical trial of intravenous diazepam and a local analgesic alone in conservative dentistry and oral surgery. J Dent Res 50:1200 (1971).
82. Bernheim J, Michiels W. Effets psychophysiques du diazepam (valium) et d'une faible dose d'alcool chez l'homme. Schweiz Med Wochenschr 103:863–870 (1973).
83. Kemp KH, Wetherell A. Some effects of a single 10 mg oral dose of diazepam on the psychomotor performance of normal men. Technical Note 326. Porton Down, UK:Chemical and Biological Defence Establishment, 1977.
84. Palva ES, Linnoila M, Saario I, Mattila MJ. Acute and subacute effects of diazepam on motor skills: interaction with alcohol. Acta Pharmacol Toxicol 45:257–264 (1979).
85. Linnoila M, Maki M. Acute effects of alcohol, diazepam, thioridazine, flupenthixole and atropine on psychomotor performance profiles. Arzneim Forsch 24:565–569 (1974).
86. Saario I, Linnoila M, Mattila MJ. Modification by diazepam or thioridazine of the psychomotor skills related to driving: a subacute trial in neurotic out-patients. Br J Clin Pharmacol 3:843–848 (1976).
87. Newman MG, Trieger N, Loskota WJ, Jacobs AW. A comparative study of psychomotor effects of intravenous agents used in dentistry. J Oral Surg 30:34–40 (1970).
88. Tansella M, Zimmermann-Tansella C, Lader M. The residual effects of N-desmethyldiazepam in patients. Psychopharmacology 38:81–90 (1974).
89. Korttila K, Linnoila M. Recovery and skills related to driving after intravenous sedation: dose-response relationship with diazepam. Br J Anaesth 41:451–463 (1975).
90. Korttila K, Linnoila M. Psychomotor skills related to driving after intramuscular administration of diazepam and meperidine. Anaesthesiol 42:685–691 (1975).
91. Hughes FW, Forney RB, Richards AB. Comparative effects in human subjects of chlordiazepoxide, diazepam and placebo on mental and physical performance. Clin Pharmacol Ther 6:139–145 (1965).
92. Linnoila M, Mattila M. Drug interaction on psychomotor skills related to driving: diazepam and alcohol. Eur J Clin Pharmacol 5:186–194 (1973).
93. Milner G, Landauer AA. Haloperidol and diazepam alone and together with alcohol in relation to driving safety. Blutalkohol 10:247–254 (1973).
94. Ogawa N, Kuwahara H, Matsuo H, Shiga K, Takata R, Agari S, Kawasaki S. An application of mirror drawing for the evaluation of diazepam in human subjects. Fukuoka Acta Med 55:915–919 (1964).
95. Linnoila M, Saario F, Maki M. Effect of treatment with diazepam or lithium and alcohol on psychomotor skills related to driving. Eur J Clin Pharmacol 7:337–342 (1974).
96. Tyrer PJ, Lader MH. Physiological and psychological effects of ± propranolol and diazepam in induced anxiety. Br J Clin Pharmacol 1:379–385 (1974).
97. Karniol IG, Dalton J, Lader MH. Comparative psychotropic effects of trazodone, imipramine and diazepam in normal subjects. Curr Ther Res 20:337–348 (1976).
98. Healy TEJ, Lautch H, Hall N, Tomlin PJ, Vickers MD. Interdisciplinary study of diazepam sedation for outpatient dentistry. Br Med J 3:13–17 (1970).
99. Hillestad L, Hanson B, Melsom H, Drivenes A. Diazepam metabolism in normal man. I. Serum concentration and clinical effect after intravenous, intramuscular and oral administration. Clin Pharmacol Ther 16:479–484 (1974).
100. Hillestad L, Hanson B, Melsom H. Diazepam metabolism in normal man. II. Serum concentration and clinical effect after oral administration and cumulation. Clin Pharmacol Ther 16:485–489 (1974).
101. Clark CH, Nicholson AN. Immediate and residual effects in man of the metabolites of diazepam. Br J Clin Pharmacol 6:325–331 (1978).
102. Cooper SA, Anthony JE, Mopsik E, Moore MS, Sullivan DC, Kruger GO. A technique for investigating the intensity and duration of human psychomotor impairment after intravenous diazepam. Oral Surg 45:493–502 (1978).
103. Savage PPE, Wilkinson V. Reaction time in psychiatric patients: a pilot study. NZ Med J 73:285–288 (1971).
104. Montagu JD. Effects of diazepam on the EEG in man. Eur J Clin Pharmacol 17:167–170 (1972).
105. Wetherell A. Effects of 5 mg diazepam on human cognitive and psychomotor performance. Technical Note 653. Chemical and Biological Defence Establishment, Porton Down, UK, 1984.
106. Ghoneim MM, Mewaldt SP, Thatcher JW. The effect of diazepam and fentanyl on mental, psychomotor and electroencephalographic functions and their rate of recovery. Psychopharmacology 44:61–66 (1975).
107. Grove-White IG, Kelman GR. Effect of methohexitone, diazepam and sodium 4-hydroxybutyrate on short-term memory. Br J Anaesth 43:113–115 (1971).
108. Frumin MJ, Herekar W, Jarvik ME. The amnesic effects of diazepam and scopolamine in man. In: Proceedings of the 4th International Congress on Pharmacology, Basle. 4:289–290 (1969).
109. George KA, Dundee JW. Relative amnesic actions of diazepam, flunitrazepam and lorazepam in man. Br J Clin Pharmacol 4:45–50 (1977).
110. Ghoneim MM, Mewaldt SP. Effects of diazepam and scopolamine on storage, retrieval and organisational processes in memory. Psychopharmacology 44:257–262 (1975).
111. Brown ID. Decrement in skill observed after seven hours of car driving. Psychon Sci 7:131–132 (1967).
112. Silverstone T. Drugs and driving. Br J Clin Pharmacol

1:451–454 (1974).
113. Naatanen R, Summala H. Road User Behaviour and Traffic Accidents. Amsterdam:North Holland, 1976.
114. O'Hanlon JF. Psychotropic Medication and Driving Safety. Rpt No VK 80-05. The Netherlands:Verkeerskundig Studiecentrum, Rijksuniversiteit Groningen, 1980.
115. Ziedman K, Smiley A, Moskowitz H. Effects of drugs on driving:driving simulator tests of diazepam and secobarbital. Proc Hum Factors 1979;259–262.
116. Dureman I, Norrman B. Clinical and experimental comparison of diazepam, chlorazepate and placebo. Psychopharmacology 40:279–284 (1975).
117. Wetherell A. Individual and group effects of 10 mg diazepam on drivers' ability, confidence and willingness to act in a gap-judging task. Psychopharmacology 63:259–267 (1979).
118. Vermeeren A, de Gier JJ, O'Hanlon JF. Methodological Guidelines for Experimental Research on Medicinal Drugs Affecting Driving Performance: an International Expert Survey. Rpt No IHP 93-27, Institute for Human Psychopharmacology, University of Limburg, Maastricht, The Netherlands, 1993.
119. Wolschrijn H, De Gier JJ, De Smet PAGM. Drugs and driving: a new categorization system for drugs affecting psychomotor performance. Instituut voor Geneesmiddelen, Veiligheid en Gedrag, University of Limburg, Maastricht, The Netherlands, 1991.
120. Guilford JP. Psychometric Methods, 2nd ed. New York:McGraw-Hill,1954.
121. Wechsler D. The Measurement of Adult Intelligence. Baltimore:Williams & Wilkins, 1944.
122. Fleishman EA, Quaintance MK. Taxonomies of Human Performance. New York:Academic Press, 1984.
123. Theologus GC, Romashko T, Fleishman EA. Development of a taxonomy of human performance: a feasibility study of ability dimensions for classifying human tasks. JSAS Cat Sel Doc Psychol 3:25–26 (1973).
124. Fleishman EA. Toward a taxonomy of human performance. Am Psychol 30:1127–1149 (1975).
125. Schemmer FM. Development of Rating Scales for Selected Visual, Auditory and Speech Abilities. Rpt No 3064. Washington:Advanced Research Resources Organization, 1982.
126. James W. Principles of Psychology. New York:Holt, 1890.
127. Boring EG. A History of Experimental Psychology. 2nd ed. New York:Appleton-Century-Crofts, 1950.
128. Broadbent DE. Perception and Communication. Oxford, UK:Pergamon Press, 1958.
129. Moray N. Where is attention limited: a survey and a model. Acta Psychol 27:84–92 (1967).
130. Kahneman D. Attention and Effort. Englewood Cliffs, NJ:Prentice-Hall, 1973.
131. Norman D, Bobrow D. On data limited and resource limited processing. J Cogni Psychol 7:44–60 (1975).
132. Navon D, Gopher D. On the economy of the human processing system. Psychol Rev 86:214–255 (1979).
133. Wickens CD. Processing resources in attention. In: Varieties of Attention (Parasuraman R, Davies DR, eds). New York:Academic Press, 1984.
134. Allport DA, Antonis B, Reynolds P. On the division of attention: a disproof of the single channel hypothesis. Q J Exp Psychol 24:225–235 (1972).
135. Treisman AM, Davies A. Divided attention to ear and eye. In: Attention and Performance IV (Kornblum S, ed). New York:Academic Press, 1973.
136. Kleiman GM. Speech recoding in reading. J Verb Learn Verb Behav 14:323–389 (1975).
137. Shaffer HL. Multiple attention in continuous verbal tasks. In: Attention and Performance V (Rabbitt PMA, Dornic S, eds). London:Academic Press, 1975.
138. Wickens CD. The effects of divided attention on information processing in manual tracking. J Exp Psychol Hum Percep Perform 2:1–13 (1976).
139. Rollins HA Hendricks R. Processing of words presented simultaneously to eye and ear. J Exp Psychol Hum Percep Perform 6:99–109 (1980).
140. Wickens CD. The structure of attentional resources. In: Attention and Performance, VIII (Nickerson R, ed). NJ:Erlbaum, Hillsdale, 1980.
141. Kantowitz BH, Knight JL. Testing tapping and time-sharing. II. Use of auditory second task. Acta Psychol 40:343–362 (1976).
142. North RA. Task Components and Demands as Factors in Dual-Task Performance. Rpt No ARL-77-2/AFOSE-77-2. Urbana, IL:Aviation Research Laboratory, 1977.
143. Wickens CD, Kessel C. The effect of participatory mode and task workload on the detection of dynamic system failures. IEEE Transact Syst Man Cybernet 13:21–31 (1979).
144. Isreal JB, Chesney GL, Wickens CD, Donchin E. P300 and tracking difficulty: evidence for multiple resources in dual-task performance. Psychophysiology 17:259–273 (1980).
145. McLeod P. A dual task response modality effect: support for multi-processor models of attention. Q J Exp Psychol 29:651–667 (1977).
146. Harris S, Owens J, North RA. A system for the assessment of human performance in concurrent verbal and manual control tasks. Behav Res Methods Instrum 10:329–333 (1978).
147. McFarland K, Ashton R. The influence of concurrent task difficulty on manual performance. Neurophysiology 16:735–741 (1978).
148. Martin M. Attention to words in different modalities: four channel presentation with physical and semantic selection. Acta Psychol 44:99–115 (1980).
149. Vidulich M, Wickens CD. Time-sharing manual control and memory search: the joint effects of input and output modality competition, priorities and control order. Technical Rpt EPL-81-4/ONR-81-4. Urbana, IL:University of Illinois, 1981.
150. Friedman A, Polson MC, Gaskill SJ, Dafoe CG. Competition for left hemisphere resources: right hemisphere superiority at abstract verbal information processing. J Exp Psychol Hum Percept Perform 7:1031–1051 (1982).
151. Wickens CD, Sandry DL. Task hemispheric integrity in dual task performance. Acta Psychol 52:227–248 (1982).
152. Wickens CD, Sandry DL, Vidulich M. Compatibility and resource competition between modalities of input, control processing and output: testing a model of complex performance. Human Factors 25: 227–248 (1983).
153. Pritchard WS, Hendrickson R. The structure of human attention: evidence for separate spatial and verbal resource pools. Bull Psychon Soc 23:177–180 (1985).
154. Kinsbourne M, Hicks R. Functional cerebral space. In: Attention and Performance VII (Requin J, ed). Hillsdale, NJ:Erlbaum, 1978.
155. Sanders AF. Some remarks on mental load. In: Mental Workload: Its Theory and Measurement (Moray N, ed). New York:Plenum Press, 1979.
156. Brooks LR. Spatial and verbal components of the act of recall. Can J Psychol 22:349–368 (1968).
157. Dimond SJ, Beaumont JG. Processing in perceptual integration between and within the cerebral hemispheres. Br J Psychol 63:509–514 (1972).
158. Allwitt LF. Two neural mechanisms related to modes of selective attention. J Exp Psychol Hum Percept Perform 7:324–332 (1981).
159. Glucksberg S. Rotary pursuit tracking with divided attention. J Eng Psychol 2:119–125 (1963).
160. Wewerwinke P. Human monitoring and control behavior. Technical Rpt NLR TR 77010. The Netherlands:National Aerospace Laboratory, 1976.
161. Lindsay PH, Taylor MM, Forbes SM. Attention and multidimensional discrimination. Percept Psychophys 4:113–117 (1968).
162. Trumbo D, Milone F. Primary task performance as a function of encoding, retention, and recall in a secondary task. J Exp Psychol 91:273–279 (1971).
163. Neisser U. Cognition and Reality. San Francisco:Freeman, 1976.

164. Reicher GM. Perceptual recognition as a function of meaningfulness of stimulus material. J Exp Psychol 81:274–280 (1969).
165. Johnson JC, McClelland JL. Perception of letters in words: seek not and ye shall find. Science 184:1192–1194 (1974).
166. Pomerantz JR, Sager LC, Stoever R. Perception of words and their component parts: some configural superiority effects. J Exp Psychol Hum Percep Perform 3:422–435 (1977).
167. Rabbitt PMA. Current paradigms and models in human information processing. In: Human Stress and Cognition (Hamilton V, Warburton DM, eds), New York:John Wiley & Sons, 1979.
168. Bainbridge L. Forgotten alternatives in skill and workload. Ergonomics 21:169–185 (1978).
169. Sperandio JC. Charge de travail et regulation des processes operatoires. Travail Humain 35:85–98 (1972).
170. Sanders AF. Towards a model of stress and human performance. Acta Psychol 53:61–97 (1983).
171. Hamilton P, Hockey GRJ, Reyman R. The place of the concept of activation in human information processing. In: Attention and Performance, VI (Dornic S, ed). Hillsdale, NJ: Erlbaum, 1977. .
172. Hockey R. Stress and the cognitive components of skilled performance. In: Human Stress and Cognition (Hamilton V, Warburton DM, eds), New York:John Wiley & Sons, 1979;141–177.
173. Gopher D, Sanders AF. "S-Oh-R": oh stages! oh resources! In: Cognition and Motor Processes (Prinz W, Sanders AF, eds). Berlin:Springer, 1984;231–253.
174. Donders FC. On the speed of mental processes. In: Attention and Performance, II (Koster WG, ed). Acta Psychol 30:412–431 (1969).
175. Sternberg S. High-speed scanning in human memory. Science 153:652–654 (1966).
176. Sternberg S. Two operations in character recognition: some evidence from reaction time measurements. Percept Psychophys 2:45–53 (1967).
177. Sternberg S. Memory scanning: mental processes revealed by reaction time experiments. Am Sci 57:421–457 (1969).
178. Sternberg S. The discovery of processing stages: extensions of Donders' method. Acta Psychol Atten Perform 30:276–315 (1969).
179. Sternberg S. Memory scanning: new findings and current controversies. Q J Exp Psychol 27:1–32 (1975).
180. Baddeley AD. The Psychology of Memory. New York:Basic Books, 1976.
181. Baddeley AD. But what the hell is it for? In: Practical Aspects of Memory, Vol 1 (Gruneberg MM, Morris PE, Sykes RN, eds). London:John Wiley & Sons, 1988;3–18.
182. Poulton EC. Range effects in experiments on people. Am J Psychol 88:3–32 (1975).
183. Millar K. The recovery of memory function. In: Ambulatory Anaesthesia and Sedation: Impairment and Recovery (Klepper ID, Sanders LD, Rosen M, eds). Oxford:Blackwell, 1991.
184. Cull C, Trimble MR. Anticonvulsant benzodiazepines and performance. Royal Soc Med Int Symposium Series 74:121–128 (1985).
185. Parrott AC. Performance tests in psychopharmacology. 1: Test reliability and standardisation. Human Psychopharmacology 6:1–9 (1991).
186. Cronbach LJ. Essentials of Psychological Testing. 4th ed. New York:Harper and Row, 1984.
187. Kelly EL. Assessment of Human Characteristics. London:Wadsworth, 1967.
188. Anastasi A. Psychological Testing. 5th ed. New York:Macmillan, 1982.
189. Kline P. A Handbook of Test Construction. London:Methuen, 1986.
190. Jones LV, Appelbaum MI. Psychometric Methods. Annu Rev Psychol 40:23–43 (1989).
191. Parrott AC. Performance tests in psychopharmacology. 2: Content validity, criterion validity and face validity. Human Psychopharmacology 6:91–98 (1991).
192. Parrott AC. Performance tests in psychopharmacology. 3: Construct validity and test interpretation. Human Psychopharmacology 6:197–207 (1991).
193. Stevens SS. Mathematics, measurement and psychophysics. In: Handbook of Experimental Psychology (Stevens SS, ed), New York:John Wiley & Sons, 1951.
194. Siegel S. Nonparametric Statistics. Tokyo:McGraw-Hill, 1956.
195. Thorndike RL. Personnel Selection. New York:John Wiley & Sons, 1949.
196. Wesnes K, Simpson P, Christmas L. The assessment of human information processing abilities in psychopharmacology. In: Human Psychopharmacology: Measures and Methods (Hindmarch I, Stonier PD, eds). Chichester, UK:John Wiley & Sons, 1987; 79–91.
197. Hockey R, Hamilton P. The cognitive patterning of stress states. In: Stress and Fatigue in Human Performance (Hockey R, ed). Chichester, UK:John Wiley & Sons, 1983;331–362.
198. Parrott AC. The effects of transdermal scopolamine and four doses of oral scopolamine (0.15, 0.3, 0.6, 1.2mg) upon psychological performance. Psychopharmacology 89:347–354 (1986).
199. Holding DH. Skills research. In: Human Skills (Holding DH, ed), Chichester, UK:John Wiley & Sons, 1989;1–16.
200 Wood CD, Manno JE, Manno BR, Redtzki HM, Wood MJ, Mims ME. Evaluation of antimotion sickness drug side effects on performance. Aviat Space Environ Med 56:310–316 (1985).
201. Henry PH, Flueck JA, Sanford JF. Assessment of performance in a Link flight simulator at three alcohol dose levels. Aerospace Med 45:33–44 (1974).
202. Billings CE, Gerke RJ, Wick RL. Comparisons of pilot performance in simulated and actual flight. Aviat Space Environ Med 46:304–308 (1975).
203. Billings CE, Wick RL, Gerke RJ, Chase RC. Effects of ethyl alcohol on pilot performance. Aerospace Med 44:379–382 (1973).
204. Seashore RH, Ivy AC. The effects of analeptic drugs in relieving fatigue. Psychol Monographs 67:1–16 (1953).
205. Hansteen RW, Miller RD, Lonero L, Reid LD, Jones B. Effects of cannabis and alcohol on automobile driving and psychomotor tracking. Ann NY Acad Sci 73:240–256 (1976).
206. De Gier JJ, 't Hart BJ, Nelemans FA, Bergman H. Psychomotor performance and real driving performance of outpatients receiving diazepam. Psychopharmacology 73:340–344 (1981).
207. Korttila K, Linnoila M. Skills related to driving after intravenous diazepam, flunitrazepam, or droperidol. Br J Anaesth 46:961–969 (1974).
208. Linnoila M. Tranquilizers and driving. Accid Anal Prev 8:15–19 (1976).
209. Linnoila M. Effect of drugs and alcohol on psychomotor skills related to driving. Ann Clin Res 6:7–18 (1974).
210. Hakkinen S. Traffic Accidents and Driving Characteristics: A Statistical and Psychological Study. Rpt No 13. Helsinki, Finland:Institute of Technology, 1958.
211. O'Hanlon JF, Haak TW, Blaauw DJ, Riemersma JBJ. Diazepam impairs lateraLondon:osition control in highway driving. Science 217:79–81 (1982).
212. Steiner-Chaskel N, Lader MH. Effects of single doses of clobazam and diazepam on psychological functions in normal subjects. Royal Soc Med Int Symp Series 43:23–32 (1981).
213. Rigal J, Savelli A. Clobazam, vigilance functions and motor car driving. Gazette Medicale de France 82:390S–391S (1975).
214. Hindmarch I, Gudgeon A. The effect of clobazam and lorazepam on aspects of psychomotor performance and car handling. Br J Clin Pharmacol 10:145–150 (1980).
215. Mackworth NH. Some factors affecting vigilance. Advancement of Science 53:389–393 (1957).
216. Smith CM. Drugs and human memory. In: Aspects of Psychopharmacology (Singer DJ, Blackman DE, eds). London:Methuen, 1984;140–173.
217. Leonard JA. 5-choice Serial Reaction Apparatus. Rpt No 326. Cambridge, UK:Medical Research Council Applied Psychology

Unit, 1959.
218. Taeuber Z, Gammel G, Gordon A, Koeppen D. Methods for the assessment of psychotropic drug effects in healthy volunteers. Mod Probl in Pharmacopsychol 12:23–26 (1977).
219. Broadbent DE. Decision and Stress. London:Academic Press, 1971.
220. Broadbent DE. Task combination and selective intake of information. Acta Psychol 50:253–290 (1982).
221. Bixler EO, Scharf MB, Leo LA, Kales A. Hypnotic drugs and performance: a review of theoretical and methodological considerations. In: Hypnotics: Methods of Development and Evaluation. New York:Spectrum, 1975;175–196.
222. Loomis TA, West TC. The influence of alcohol on automobile driving ability. Q J Stud Alcohol 19:30–46 (1958).
223. Drew GC, Colquhoun WP, Long HA. Effects of small doses of alcohol on a skill resembling driving. Medical Research Council Memo 38, London:Her Majesty's Stationery Office, 1959.
224. Mortimer RG. Effect of low blood-alcohol concentrations in simulated day and night driving. Percept Mot Skills 17:399–408 (1963).
225. Light WO, Keiper CG. Effects of Moderate Blood-Alcohol Levels on Automobile Passing Behavior. Rpt No ICRL-RR-69-4. Washington:Department of Health, Education and Welfare, 1969.
226. Crancer A, Dille JM, Delay JC, Wallace JE, Haykin MD. Comparison of the effects of marihuana and alcohol on simulated driving performance. Science 164:851–854 (1969).
227. Rafaelson OJ, Bech P, Christiansen J, Christup H, Nyboe J, Rafaelson L. Cannabis and alcohol: effects on simulated car driving. Science 179:920–923 (1973).
228. Landauer AA, Milner G, Patman J. Alcohol and amitryptiline effects on skills related to driving behavior. Science 163:1467–1468 (1969).
229. Patman J, Landauer AA, Milner G. The combined effects of alcohol and amitryptiline on skills similar to motor-car driving. Med J Aust 2:946–949 (1969).
230. Hughes DTD, Cramer F, Knight GJ. Use of a racing car simulator for medical research: the effects of marzine and alcohol on driving performance. Med Sci Law 7:200–204 (1967).
231. Milner G, Landauer AA. Alcohol, thioridazine and chlorpromazine effects on skills related to driving. Br J Psychiat 118:351–352 (1971).
232. Goldman V, Comerford B, Hughes D, Nyberg G. Effect of beta-adrenergic blockade and alcohol on simulated car driving. Nature 224:1175–1178 (1969).
233. Dott AB. Effect of Marihuana on Risk Acceptance in a Simulated Passing Task. Public Health Service Rpt ICRL-RR-71-3. Washington:Department of Health, Education and Welfare, 1972.
234. Green R, Long HA, Eliott CJR, Howells TH. A method of studying recovery after anaesthesia. Anaesthesia 18:189–200 (1963).
235. Wilkinson BM. Driving ability and reaction times following intravenous anaesthesia. NZ Dent J 61:21–26 (1965).
236. Moore NC. Medazepam and the driving ability of anxious patients. Psychopharmacology 52:103–106 (1977).
237. Uhr L, Pollard JC, Miller JG. Behavioral effect of chronic administration of psychoactive drugs to anxious patients. Psychopharmacology 1:150–168 (1959).
238. Miller JG. Objective measurement of the effect of drugs on driver behavior. J Am Med Assoc 179:940–943 (1962).
239. Ashton H, Savage RD, Telford R, Thompson JW, Watson DW. The effects of cigarette smoking on the response to stress in a driving simulator. Br J Clin Pharmacol 45:546–556 (1972).
240. Klonoff H. Marihuana and driving in real-life situations. Science 186:317–324 (1974).
241. Kielholz P, Goldberg L, Im Obersteg I, Poldinger W, Ramseyer A, Schmidt P. Strassenverkehr, Tranquilizer und Alkohol. Deutsche Med Wochensch 92:1525–1531 (1967).
242. Kielholz P, Goldberg L, Im Obersteg I, Ramseyer A, Schmidt P. Fahrversuche zur Frage der Beeträchtigung der Verkehrstuchtigkeit durch Alkohol, Tranquilizer und Hypnotika. Deutsche Med Wochensch 94:301–306 (1969).
243. Betts TA, Clayton AB, Mackay GM. Effects of four commonly used tranquillisers on low-speed driving performance tests. Br Med J 4:580–584 (1972).
244. Hindmarch I, Hanks GW, Hewett AJ. Clobazam, a 1,5-benzodiazepine, and car driving ability. Br J Clin Pharmacol 4:573–578 (1977).
245. Clayton AB, Harvey PC, Betts TA. The effects of two antidepressants, imipramine and viloxazine, upon driving performance. Psychopharmacology 55:9–12 (1977).
246. Smiley A, LeBlanc E, French J, Burford R. The combined effects of alcohol and common psychoactive drugs: field studies with an instrumented automobile. Can Soc Forensic Sci 8:57–64 (1975).
247. Wetherell A. Effects of atropine on drivers' perceptual motor and decision making behaviour. In: Drugs and Driving (O'Hanlon JF, de Gier JJ, eds). London:Taylor & Francis, 1986.
248. Wetherell A. Drugs and drivers' visual perception. In: Vision in Vehicles (Gale AG, Freeman MH, Smith P, Taylor SP, eds). Amsterdam:North Holland, 1986.
249. Cohen J, Dearnaley EJ, Hansel CEM. The risk taken in driving under the influence of alcohol. Br Med J 2:1438–1442 (1958).
250. Brown ID, Tickner AH, Simmonds DCV. Interference between concurrent tasks of driving and telephoning. J Appl Psychol 53:419–424 (1969).
251. Barrett GV, Alexander RA, Forbes JB. Analysis of Performance Measurement and Training Requirements for Drivers Decision Making in Emergency Situations. Rpt No DOT HS-800 867. New York:Management Research Center, University of Rochester, 1973.
252. Harms PL. Driver Following Studies on the M4 Motorway During a Holiday and a Normal Weekend in 1966. Rpt No LR136. Crowthorne, UK:Transport and Road Research Laboratory, 1966.
253. Colbourn CJ, Brown ID, Copeman AK. Drivers' judgments of safe distances in vehicle following. Hum Factors 20:1–11 (1978).
254. Rockwell TH, Snider JN. An Investigation of Variability in Driving Performance. Rpt No RF 1450. Columbus, OH:Research Foundation, Ohio State University, 1967.
255. De Gier JJ, Kuijpens L, Nelemans FA. The effects of astemizole on actual car driving and psychomotor performance. In: Drugs and Driving (O'Hanlon JF, De Gier JJ, eds). London:Taylor & Francis, 1986;271–282.
256. Wilkinson RT, Houghton D. Portable four choice reaction time test with magnetic tape memory. Behav Res Methods Instrum 7:441–446 (1975).
257. Wesnes K, Warburton DM. Effects of smoking on rapid information processing performance. Neuropsychobiology 9:223–229 (1983).
258. Wesnes K, Warburton DM. Effects of scopolamine and nicotine on human rapid information processing performance. Psychopharmacology 82:147–150 (1984).
259. AGARD. Human Performance Assessment Methods. AGARDograph 308. Neuilly-sur-Seine, France:NATO, Advisory Group for Aerospace Research and Development, 1989.
260. AGARD. Human Performance Assessment Methods. AGARDograph 308 Addendum. Neuilly-sur-Seine, France:NATO, Advisory Group for Aerospace Research and Development, 1991.
261. Wetherell A. The STRES Battery: standardised tests for research into environmental stress. In: Contemporary Ergonomics, 1990 (Lovesey EJ, ed). London:Taylor & Francis, 1990;270–275.
262. Boer LC, Gaillard AWK, Jorna PGAM. Taskomat: a task battery for information processing [in Dutch]. Rpt IZF 1987-2. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie TNO, 1987.
263. Moraal J. Age and Information Processing: an Application of

Sternberg's Additive Factor Method. Rpt No IZF 1982-18. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie TNO, 1982.
264. Gaillard AWK, Verduin CJ. The combined effects of an antihistamine and pseudo-ephedrine on human performance. J Drug Res 8:1929–1936 (1983).
265. Gaillard AWK, Gruisen A, de Jong R. The Influence of Loratidine (Sch 29851) on Human Performance. Rpt No IZF 1986-C19. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie, 1986.
266. Gaillard AWK, Rozendaal AH, Varey CA. Marginal Vitamin Deficiency Aand Mental Performance [in Dutch]. Rpt No IZF 1983-29. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie, 1983.
267. Gaillard AWK, Varey CA, Ruzius MHB. Marginal Vitamin Deficiency and Mental Performance. Rpt No IZF 1985-22. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie, 1985.
268. Boer LC, Ruzius MHB, Mimpen AM, Bles W, Janssen WH. Psychological Fitness during a Manoeuvre. Rpt No IZF 1984-17. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie, 1984.
269. Stokx LC, Gaillard AWK. Task and driving performance of patients with a severe concussion of the brain. J Clin Exp Neuropsychol 8:421–436 (1986).
270. Sanders AF, Wijnen JLC, van Arkel AE. An additive factor analysis of the effects of sleep loss on reaction processes. Acta Psychol 51:41–59 (1982).
271. Steyvers FJJM. The influence of sleep deprivation and knowledge of results on perceptual encoding. Acta Psychol 66:173–178 (1987).
272. Frowein HW, Gaillard AWK, Varey CA. EP components, visual processing stages, and the effect of a barbiturate. Biol Psychol 13:239–249 (1981).
273. Logsdon R, Hochhaus L, Williams HL, Rundell OH, Maxwell D. Secobarbital and perceptual processing. Acta Psychol 55:179–193 (1984).
274. Frowein HW, Reitsma D, Aquarius C. Effects of two counteracting stressors on the reaction process. In: Attention and Performance, IX (Long J, Baddeley AD, eds). Hillsdale, NJ:Erlbaum, 1981.
275. Moran LJ, Mefferd RB. Repetitive psychometric measures. Psychol Rep 5:269–275 (1959).
276. Hart JJ. Standardisation studies with the Repetitive Psychometric Measures I: determining equivalence of forms. Technical Memo 2–17. Edgewood, MD:Edgewood Arsenal, 1965.
277. Hart JJ, Kysor KP. Standardisation studies with the Repetitive Psychometric Measures III: determining the effects of ability and practise on level of performance. Technical Memo 114–1, Edgewood, MD:Edgewood Arsenal, 1966.
278. Kysor KP. Standardisation studies with the Repetitive Psychometric Measures II: a comparison with the zero input tracking analyzer and an anagram test. Technical Memo 114–13, Edgewood, MD:Edgewood Arsenal, 1967.
279. Kysor KP, Hart JJ. The practise effect on number facility performance. Technical Rpt 4245, Edgewood, MD:1969.
280. Wetherell A. Practice effects on the number facility test and some methods for their control. Technical Note 347. Porton Down, UK:Chemical and Biological Defence Establishment, 1978.
281. Wetherell A. A cheap microcomputer-based psychological performance test system. Technical Note 721. Porton Down, UK:Chemical and Biological Defence Establishment, 1985.
282. Chiles WD, Alluisi EA, Adams OS. Work schedules and performance during confinement. Hum Factors 10:143–196 (1968).
283. Hall TJ, Passey GE, Meighan TW. Performance of Vigilance and Monitoring Tasks as a Function of Workload. Aerospace Medical Research Laboratory:Rpt No AMRL-TR-65-22, Wright-Patterson Air Force Base, OH, 1965.
284. Chiles WD, Bruni CB, Lewis RA. Methodology in the Assessment Of Complex Human Performance: The Effects of Signal Rate on Monitoring a Dynamic Process. Federal Aviation Administration:Rpt No FAA-AM-69-9. Washington: Department of Transportation, 1969.
285. Chiles WD, Alluisi EA. On the specification of operator or occupational workload with performance measurement methods. Hum Factors 21:515–528 (1979).
286. Wanner E, Shiner S. Measuring transient memory load. J Verb Learn Verb Behav 15:159–167 (1976).
287. Perez WA. Mental Arithmetic: the Processing and Maintenance of Information in Working Memory. PhD dissertation. Miami University, Miami, OH, 1982.
288. Shingledecker CA. A task battery for applied human performance assessment research. Aerospace Medical Research Laboratory:Rpt No AFAMRL-TR-84-071. Wright-Patterson Air Force Base, OH,1984.
289. Moran LJ, Mefferd RB, Kimble JP. Repetitive psychometric measures: equating alternate forms. Psychol Rep 47:243–251 (1964).
290. Seales DM, Kennedy RS, Bittner AC. Development of performance evaluation tests for environmental research (PETER): arithmetic computation. Percept Mot Skills 51:1023–1031 (1980).
291. Chiles WD, Jennings AE, Alluisi EA. The Measurement and Scaling of Workload in Complex Performance. Rpt No FAA-AM-7834. Washington:Federal Aviation Administration, Department of Transportation, 1978.
292. Ashcraft MH, Battaglia J. Cognitive arithmetic: evidence for retrieval and decision processes in mental addition. J Exp Psychol Hum Learn Mem 4:527–538 (1978).
293. Ashcraft MH, Stazyk EH. Mental addition: a test of three verification models. Mem Cog 9:185–196 (1981).
294. Stazyk EH, Ashcraft MH, Hamann MS. A network approach to mental multiplication. J Exp Psychol Learn Mem Cog 8:320–335 (1982).
295. Hitch GJ. The role of short term working memory in mental arithmetic. Cog Psychol 10:302–323 (1978).
296. Rauch TM, Tharion WJ The Effects of Wearing the Chemical Protective Mask and Gloves on Cognitive Problem Solving. Technical Rpt T20-87. Natick, MA:US Army Research Institute of Environmental Medicine, 1987.
297. Rauch TM, Witt C, Banderet LE, Tauson R, Golden M. The Effects of Wearing Chemical Protective Clothing on Cognitive Problem Solving. Technical Rpt T18-86. Natick, MA:US Army Research Institute of Environmental Medicine, 1986.
298. Wetherell A. Effects of wearing the S6 respirator for 6 hours on the cognitive and psychomotor performance of male and female subjects. Technical Note 989, Porton Down, UK:Chemical and Biological Defence Establishment, 1989.
299. Shattock JA, Wetherell A. Cognitive, psychomotor and subjective effects of wearing full NBC Individual protective equipment for six hours. Technical Note 709, Porton Down, UK:Chemical and Biological Defence Establishment, 1994.
300. Chiles WD, Jennings AE. Effects of alcohol on complex performance. Hum Factors 12:605–612 (1970).
301. Kemp KH, Wetherell A. Some effects of intramuscular injections of 2 mg atropine sulphate and 5 mg diazepam on human cognitive and psychomotor performance. Technical Note 532, Chemical and Biological Defence Establishment, Porton Down, UK:1982.
302. Kemp KH, Wetherell A. Psychomotor and cognitive effects of the nerve agent immediate drug treatment: 2 mg atropine sulphate and 500 mg pralidoxime mesylate (P2S) intramuscularly, with 5 and 10 mg diazepam orally. Technical Note 547, Porton Down, UK:Chemical and Biological Defence Establishment, 1982.
303. Holland P, Kemp KH, Wetherell A. Some effects of 2 mg atropine and 5 mg diazepam, separately and combined, on human performance. Br J Clin Pharmacol 5:367P (1978).
304. Kemp KH, Wetherell A. A laboratory study of the performance of men taking pyridostigmine bromide orally (30 mg eight hourly) for two weeks. Technical Note 494, Porton Down,

UK:Chemical and Biological Defence Establishment, 1981.
305. Doherty PC, Wetherell A. Cognitive, psychomotor and subjective effects of 0.6 mg hyoscine. Technical Note 1055, Porton Down, UK:Chemical and Biological Defence Establishment, 1990.
306. Toulmin SJ, Wetherell A. Some effects of anticholinergic drugs on performance. In: Contemporary Ergonomics 1995 (Robertson, SA ed). London:Taylor & Francis, 1995; 505–510.
307. Schlegel RE, Gilliland K. Evaluation of the Criterion Task Set. Rpt No AAMRL-TR-87. Wright-Patterson Air Force Base, OH:Armstrong Aerospace Medical Research Laboratory, 1989.
308. Repko JD, Jones PE, Garcia LS, Schneider EJ, Roseman E, Corum CR. Behavioral and neurological effects of methyl chloride. (NIOSH) Publ No 77-125, Cincinnati, OH:National Institute for Occupational safety and Health, 1976.
309. Graboi D. Searching for targets: the effects of specific practice. Percept Psychophys 10:300–304 (1971).
310. Klatzky RL, Juola JF, Atkinson RC. Test stimulus presentation and experimental context effects in memory scanning. J Exp Psychol 87:281–288 (1971).
311. Lively BL, Sanford BJ. The use of category information in a memory search task. J Exp Psychol 93:379–385 (1972).
312. Naus MJ, Glucksberg S, Ornstein PA. Taxonomic word categories and memory search. Cog Psychol 3:643–654 (1972).
313. Burrows D, Okada R. Serial position effects in high-speed memory search. Percept Psychophys 10:305–308 (1971).
314. Connor JM. Serial and parallel encoding processes in memory and visual search. J Exp Psychol 96:363–370 (1972).
315. DeRosa DV, Morin RE. Recognition and reaction time for digits in consecutive and nonconsecutive sets. J Exp Psychol 83:472–479 (1970).
316. Morin RE, DeRosa DV, Stultz V. Recognition memory and reaction time. Acta Psychol 27:298–305 (1967).
317. Lively BL. Speed/accuracy tradeoff and practice as determinants of stage duration in a memory search task. J Exp Psychol 96:97–103 (1972).
318. DeRosa DV. Transformation on sets in short term memory: set size reduction by deletion. J Exp Psychol 82:415–426 (1969).
319. DeRosa DV, Sabol M. Transformation on sets in short term memory: temporal and spatial factors influencing deletion. Mem Cog 1:69–72 (1973).
320. Briggs GE, Blaha J. Memory retrieval and central comparison times in information processing. J Exp Psychol 79:395–402 (1969).
321. Briggs GE, Johnsen AM. On the nature of central processing in choice reactions. Mem Cog 1:91–100 (1973).
322. Briggs GE, Swanson JM. Retrieval time as a function of memory ensemble size. Q J Exp Psychol 21:185–191 (1969).
323. Klatzky RL, Atkinson RC. Specialization of the hemispheres in scanning for information in short term memory. Percept Psychophys 10:335–338 (1971).
324. Chase WG, Calfee RC. Modality and similarity effects in short term recognition memory. J Exp Psychol 81:510–514 (1969).
325. Hoving KL, Morin RE, Konick DS. Recognition reaction time and size of the memory set: a developmental study. Psychon Sci 21:247–248 (1970).
326. Foss DJ, Dowell BE. High-speed memory retrieval with auditorily presented stimuli. Percept Psychophys 9:465–468 (1971).
327. Swanson JM, Johnsen AM, Briggs GE. Recoding in a memory search task. J Exp Psychol 93:1–9 (1972).
328. Burrows D, Okada D. Parallel scanning of semantic and formal information. J Exp Psychol 97:254–257 (1973).
329. Clifton C, Tash J. Effect of syllabic word length on memory search rate. J Exp Psychol 99:231–235 (1973).
330. Orenstein HB, Hamilton KM. Memory load, critical features and retrieval processes in facial recognition. Percept Mot Skills 45:1079–1087 (1977).
331. Wingfield A, Branca AA. Strategy in high-speed memory search. J Exp Psychol 83:63–67 (1970).
332. Naus MJ. Memory search of categorized lists: a consideration of alternative self-terminating search strategies. J Exp Psychol 102:992–1000 (1974).
333. Anders TR, Fozard JL, Lillyquist TD. Effects of age upon retrieval from short term memory. Dev Psychol 6:214–217 (1972).
334. Harris GJ, Fleer RE. High-speed memory scanning in mental retardates: evidence for a central processing deficit. J Exp Child Psychol 17:452–459 (1974).
335. Eysenck MW, Eysenck MC. Memory scanning, introversion-extraversion and levels of processing. J Pers Res 13:305–315 (1979).
336. Ross J. Extended practice with a single-character classification task. Percept Psychophys 8:276–278 (1970).
337. Kristofferson MW. When item recognition and visual search functions are similar. Percept Psychophys 12:379–384 (1972).
338. Kristofferson MW. Effects of practice on character classification performance. Can J Psychol 26:540–560 (1972).
339. Smith PJ, Langolf GD. The use of Sternberg's memory scanning paradigm in assessing the effects of chemical exposure. Hum Factors 23:701–708 (1981).
340. Maizlish NA, Langolf GD, Whitehead LW, Fine LJ, Albers JW, Goldberg J, Smith P. Behavioural evaluation of workers exposed to mixtures of organic solvents. Br J Indus Med 42:579–590 (1985).
341. Oborne DJ, Rogers Y. Interactions of alcohol and caffeine on human reaction time. Aviat Space Environ Med 54:528–534 (1983).
342. Tharp VK, Rundell OH, Lester BK, Williams HL. Alcohol and information processing. Psychopharmacology 40:33–52 (1974).
343. Subhan Z. The effects of benzodiazepines on short term memory and information processing. Psychopharmacology 1:173–181 (1984).
344. Rizzuto AP. Diazepam and Its Effects on Psychophysiological and Behavioral Measures of Performance. Rpt No AAMRL-TR-87-074. Wright-Patterson Air Force Base, OH:Armstrong Aerospace Medical Research Laboratory, 1987.
345. Rundell OH, Williams HL, Lester BK. Secobarbital and information processing. Percept Mot Skills 46:1255–1264 (1978).
346. Williams HL, Rundell OH, Smith LT. Dose effects of secobarbital in a Sternberg memory scanning task. Psychopharmacology 72:161–165 (1981).
347. Mohs RC, Tinklenberg JR, Roth WY, Kopell BS. Sensitivity of some human cognitive functions to effects of methamphetamine and secobarbital. Drug Alcohol Depend 5:145–150 (1980).
348. McNair DM, Kahn RJ, Frankenthaler LM, Faldetta LL. Amoxapine and amitryptiline. II: Specificity of cognitive effects during brief treatment of depression. Psychopharmacology 83:134–139 (1984).
349. Naylor H, Halliday R, Callaway E. The effect of methylphenidate on information processing. Psychopharmacology 86:90–95 (1985).
350. Wetherell A. Effects of physostigmine on stimulus encoding in a memory scanning task. Psychopharmacology 109:198–202 (1992).
351 Ward MM, Sandman CA, George JM, Shulman H. MSH Melanocyte stimulating hormone ACTH 4-10 in men and women: effects upon performance of an attention and memory task. Physiol Behav 22:669–674 (1979).
352. Briggs GE, Peters GL, Fisher RP. On the locus of the divided attention effects. Percept Psychophys 11:315–320 (1972).
353. Crosby JV, Parkinson GR. A dual task investigation of pilots' skill level. Ergonomics 22:1301–1313 (1979).
354. Wetherell A. The efficacy of some auditory-vocal subsidiary tasks as measures of the mental load on male and female drivers. Ergonomics 24:197–214 (1981).
355. Gomer FE, Spicuzza RJ, O'Donnell RD. Evoked potential correlates of visual item recognition during memory scanning tasks. Physiol Psychol 4:61–65 (1976).
356. Brookhuis KA, Mulder G, Mulder LJM, Gloerich ABM, van Dellen HJ, van der Meere JJ, Ellerman H. Late positive components and stimulus evaluation time. Biol Psychiat

13:107–123 (1981).
357. Adam N, Collins G. Late components of the visual evoked potential to search in short term memory. Electroencephalogr Clin Neurophysiol 44:147–156 (1978).
358. Ford JM, Roth WT, Mohs RC, Hopkins WF, Kopell BS. Event related potentials recorded from young and old adults during a memory retrieval task. Electroencephalogr Clin Neurophysiol 47:450–459 (1979).
359. Pfefferbaum A, Ford JM, Roth WT, Kopell BS. Age differences in P300 reaction time associations. Electroencephalogr Clin Neurophysiol 49:257–265 (1980).
360. Carter RC, Kennedy RS, Bittner AC, Krause M. Item recognition as a performance evaluation tests for environmental research. In: Proceedings of the 24th Annual Meeting Human Factors Society, Santa Monica, California, 1980;340–344.
361. Carter RC, Krause M. Reliability of Slope Scores for Individuals: Experiment 4: Letter Search. Rpt No NBDL-83R003. New Orleans:Naval Dynamics Laboratory, 1983
362 Boer LC. A Normative Database for, and Analysis of, Some Taskomat Tasks. Rpt No IZF 1988-10. Soesterberg, The Netherlands:Instituut voor Zintuigfysiologie, 1988.
363. McFarlane MA. A study of practical ability. Br J Psychol Monogr 8, 1925.
364. Thurstone LL. Primary mental abilities. Psychom Monogr 1, 1938.
365. Cattell RB. Some theoretical issues in adult intelligence testing. Psychol Bull 38:592 (1941).
366. Cattell RB. Theory of fluid and crystallized intelligence: a critical experiment. J Ed Psychol 54:1–22 (1963).
367. Pawlik K. Faktorenanalytische Persönlichkeitsforschung. In: Kindlers Enzyklopädie der Psychologie des 20 Jahrhunderts. Zurich, 1973.
368. Dunnette MD. Aptitudes, abilities and skills. In: Handbook of Industrial and Organizational Psychology (Dunnette MD, ed). Chicago:Rand McNally, 1976.
369. Guilford JP. Way Beyond the IQ. Buffalo, NY:Creative Education Foundation, 1977.
370. Lohman DF. Spatial Ability: A Review and Reanalysis of the Correlational Literature. Technical Rpt No 8. Stanford, CA:Stanford University, 1979.
371. Shepard RN, Metzler J. Mental rotation of three-dimensional objects. Science 171:701–703 (1971).
372. Cooper LA, Shepard LN. Chronometric studies of the rotation of mental images. In: Visual Information Processing (Chase WG, ed). New York:Academic Press, 1973.
373. Egan DE. Characterizing Spatial Ability: Different Mental Processes Reflected in Accuracy and Latency Scores. Rpt No 1224. Pensacola, FL:Naval Aerospace Medical Research Laboratory, 1978.
374. Fitts PM, Weinstein M, Rappaport M, Anderson N, Leonard JA. Stimulus correlates of visual pattern perception: a probability approach. J Exp Psychol 51:1–11 (1956).
375. Kennedy RS, Dunlap WP, Jones MB, Lane NE, Wilkes RL. Portable Human Assessment Battery: Stability, Reliability, Factor Structure, and Correlation With Tests of Intelligence. Final Technical Rpt. Washington:National Science Foundation, 1985.
376. Klein R, Armitage R. Rhythms in human performance:1 1/2-hour oscillations in cognitive style. Science 204:1326–1328 (1979).
377. Lewis VJ, Baddeley AD. Cognitive performance, sleep quality, and mood during deep oxyhelium diving. Ergonomics 24:773–793 (1981).
378. Logie RH, Baddeley AD. A trimix saturation dive to 660 m: studies of cognitive performance, mood, and sleep quality. Ergonomics 26:359–374 (1983).
379. Wickens CD. Engineering Psychology and Human Performance. Columbus, OH:Charles E Merrill, 1984.
380. Ashkenas IL, McRuer DT. The Determination of Material Handling Quality Requirements from Airframe/Human-pilot System Studies. Rpt TR-59-135. Wright-Patterson Air Force Base, OH:Wright Air Development Center, 1959.
381. Jex HR, Cromwell CH. Theoretical and Experimental Investigation of Some New Longitudinal Handling Quality Parameters. Rpt TR 61-26. Wright-Patterson Air Force Base, OH:Aeronautical Systems Division, 1961.
382. Jex HR, McDonnell JD, Phatak AV. A "critical" tracking task for manual control research. IEEE Transac Hum Factors Electron 7:138–144 (1966).
383. McRuer DT, Jex HR. A review of quasilinear pilot models. IEEE Transac Hum Factors Electron 8:231–249 (1967).
384. Damos DL, Bittner AC, Kennedy RS, Harbeson MM. Effects of extended practice on dual task tracking performance. Hum Factors 23:627–631 (1981).
385. Damos DL, Bittner AC, Kennedy RS, Harbeson MM, Krause MK. Performance evaluation tests for environmental research (PETER): critical tracking test. Percept Mot Skills 58:567–573 (1984).
386. Klein R, Jex HR. Effects of alcohol on a critical tracking task. J Studies on Alcohol 36:11–20 (1975)
387. Dott AB, McKelvy RK. Influence of ethyl alcohol in moderate levels on visual stimulus tracking. Hum Factors 19:191–199 (1977).
388. Putz VR. The effects of carbon monoxide on dual task performance. Hum Factors 21:13–24 (1975).
389. Grether WF. Acceleration and Human Performance. R p t AMRL-TR-71-22. Wright-Patterson Air Force Base, OH:Aerospace Medical Research Laboratory, 1971.
390. Jex HR, Peters RA, DiMarco RJ, Allen RW. The Effects of Bedrest on Crew Performance during Simulated Shuttle Reentry. Vol II: Control task performance. Rpt No NASA CR 2367. Washington:National Aeronautics and Space Administration, 1974.
391. Van Patten RE. Tolerance, fatigue, physiological and performance effects of sustained and oscillating lateral acceleration. In: The Conference Proceedings of the Advisory Group for Aerospace Research and Development, No 371: Human Factors Considerations in High Performance Aircraft, 1984. Paris:AGARD, 1984.
392. Wason PC. Response to affirmative and negative binary statements. Br J Psychol 52:133–142 (1961).
393. Slobin DI. Grammatical transformations and sentence comprehension in childhood and adulthood. J Verb Learn Verb Behav 5:219–227 (1966).
394. Chase WG, Clark HH. Mental operations in the comparison of sentences and pictures. In: Cognition in Learning and Memory. (Gregg W, ed). New York:John Wiley & Sons, 1972.
395 Clark HH, Chase WG. On the process of comparing sentences against pictures. Cog Psychol 3:472–517 (1972).
396. Clark HH, Chase WG. Perceptual coding strategies in the formation and verification of descriptions. Mem Cog 2:101–111 (1974).
397. Carter RC, Kennedy RS, Bittner AC. Grammatical reasoning: a stable performance yardstick. Hum Factors 23:587–591 (1981).
398. Wetherell A. Studies on a test of higher mental function. Technical Note 278, Porton Down, UK:Chemical and Biological Defence Establishment, 1976.
399. Wetherell A. Some factors affecting spatial memory for route information. In: Information Design. (Easterby RS, Zwaga HJG, eds). Chichester, UK:John Wiley & Sons, 1984.
400. Farmer EW, Berman JVF, Fletcher YL. Evidence for a visuo-spatial scratch-pad in working memory. Q J Exp Psychol 38A:675–688 (1986).
401. Baddeley AD, Hitch GJ. Working memory. In: The Psychology of Learning and Motivation. Vol 8 (Bower GH, ed). New York:Academic Press, 1974.
402. Hitch GJ, Baddeley AD. Verbal reasoning and working memory. Q J Exp Psychol 28:603–621 (1976).
403. Wetherell A. Physostigmine and memory. Technical Note 634, Porton Down, UK:Chemical and Biological Defence Establishment, 1984.
404. Baddeley AD, de Figueredo JW, Hawkswell Curtis JW, Williams AN. Nitrogen narcosis and underwater performance.

403. Wetherell A. Physostigmine and memory. Technical Note 634, Porton Down, UK:Chemical and Biological Defence Establishment, 1984.
404. Baddeley AD, de Figueredo JW, Hawkswell Curtis JW, Williams AN. Nitrogen narcosis and underwater performance. Ergonomics 11:157–164 (1968).
405. Naitoh P, Angus RG. Napping and human functioning during prolonged work. In: Sleep and Alertness: Chronobiological Behavioral and Medical Aspects of Napping. (Dinges DF, Broughton RD, eds). New York:Raven Press, 1989.
406. Ussher MH, Farmer EW. Anxiety prior to and during decompression. In: Contemporary Ergonomics 1987 (Megaw ED, ed), London:Taylor & Francis, 1987.
407. Salame P. The AGARD grammatical reasoning task: a defect and proposed solutions. Ergonomics 36:1457–1464 (1993).
408. Wickens CD, Mountford SJ, Schreiner W. Task Dependent Differences and Individual Differences in Dual Task Performance. Rpt No NBDL-M003. New Orleans:Naval Biodynamics Laboratory, 1980.
409. Sverko B. Individual Differences in Time-Sharing Performance. Savoy Aviation Research Laboratory:Rpt No ARL-77-4/ AFOSR -77-4. Champagne, IL:University of Illinois, 1977.
410. Keele SW, Hawkins HL. Explorations of individual differences relevant to high skill level. J Mot Behav 14:3–23 (1982).
411. Putz VR, Anderson V, Setzer JV, Croxton JS. Effects of alcohol, caffeine and methyl chloride on man. Psychol Rep 48:715–725 (1981).
412. Putz VR, Johnson BL, Setzer JV. A comparative study of the effects of carbon monoxide and methylene chloride on human performance. Pathol Toxicol 2:97–112 (1979).
413. Houghton JO, McBride DK, Hannah K. Performance and physiological effects of acceleration-induced (+Gz) loss of consciousness. Aviat Space Environ Med 56:956–965 (1985).
414. Farmer EW, Green RG. The sleep-deprived pilot: performance and EEG response. In: Report of the 16th Conference of the Western European Association for Aviation Psychology (Sorsa M, ed), Finnair Training Centre, Helsinki, 1985. Helsinki:Finnair, 1985; 155–162.
415. Poulton EC, Freeman PR. Unwanted asymmetrical transfer effects with balanced experimental designs. Psychol Bull 66:1–8 (1966).
416. Rabbitt PMA. The faster the better? Some comments on the use of information processing rate as an index of change and individual difference in performance. In: Psychopharmacology and Reaction Time (Hindmarch I, Aufdembrinke B, Ott H, eds), Chichester, UK:John Wiley & Sons, 1988;79–96.
417. Mayer SE, Bain JA. Localisation of the hematoencephalic barrier with fluorescent quaternary acridones. J Pharmacol Exp Ther 118:17–25 (1956).
418. Rapoport SI. Blood brain barrier in physiology and medicine. New York:Raven Press, 1976.